NOTICE: this is the author's version of a work that was accepted for publication in *Acta Psychologica*. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in 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(2), 162-174. doi:10.1016/j.actpsy.2010.01.007

<http://www.sciencedirect.com/science/article/pii/S0001691810000168>

**Preferences for different questions when testing hypotheses in an abstract task:
positivity does play a role, asymmetry does not**

Paolo Cherubini ^a, Patrice Rusconi ^a, Selena Russo ^b,

Selenia Di Bari^a, Simona Sacchi ^a

^a Department of Psychology, University of Milan-Bicocca, Italy

^b Department of Cognitive Sciences and Education, University of Trento, Italy

Author for correspondence is P.C.:

Department of Psychology

Università Milano - Bicocca

1, Piazza Ateneo Nuovo

20126, Milano, Italy

e-mail: paolo.cherubini@unimib.it

fax: ++39 02 6448 3706

tel.: ++39 02 6448 3811

Abstract

Previous studies on hypothesis-testing behaviour have reported systematic preferences for posing positive questions (i.e., inquiries about features that are consistent with the truth of the hypothesis) and different types of asymmetric questions (i.e., questions where the hypothesis confirming and the hypothesis disconfirming responses have different evidential strength). Both tendencies can contribute – in some circumstances – to confirmation biases (i.e., the improper acceptance or maintenance of an incorrect hypothesis). The empirical support for asymmetric testing is, however, scarce and partly contradictory, and the relative strength of positive testing and asymmetric testing has not been empirically compared. In four studies where subjects were asked to select (Experiment 1) or evaluate (Experiment 2, 3, 4) questions for controlling an abstract hypothesis, we orthogonally balanced the positivity/negativity of questions by their symmetry/asymmetry (Experiment 1, 2, 3), or by the type of asymmetry (confirmatory vs disconfirmatory; Experiment 4). In all Experiments participants strongly preferred positive to negative questions. Their choices were on the other hand mostly unaffected by symmetry and asymmetry in general, or – more specifically – by different types of asymmetry. Other results indicated that participants were sensitive to the diagnosticity of the questions (Experiments 1,2,3), and that they preferred testing features with a high probability under the focal hypothesis (Experiment 4). In the discussion we argue that recourse to asymmetric testing – observed in some previous studies using more contextualized problems – probably depends on context-related motivations and prior-knowledge. In abstract tasks, where that knowledge is not available, more simple strategies – such as positive testing – are prevalent.

PsycINFO classification: 2340 (Cognitive Processes)

Keywords: hypothesis testing; information selection; probabilistic reasoning.

Whenever we explicitly test the plausibility of a hypothesis, we ask questions either to ourselves, or to other people and external data bases. Since gathering all the evidence needed for an exhaustive check is seldom feasible, giving priority to some questions implies giving priority to some pieces of information above others. Different studies have emphasized that some human trends in gathering information might – in certain environments – cause undesirable side effects such as confirmation biases (Wason, 1960, 1968; Klayman & Ha, 1987; Nickerson, 1996, 1998) or the preservation of social stereotypes (Trope & Thompson, 1997; Cameron & Trope, 2004). The main goal of this study is to further explore two of those tendencies, namely the alleged preference for posing *asymmetrical questions* and *positive questions* (the latter, also known as “congruent” questions, Baron, Beattie, & Hershey, 1988), in order to verify their actual occurrence and to compare their relative strengths in abstract tasks, where domain-specific motivations and prior knowledge are hardly accessible.

1. Positivity and Asymmetry of Questions

A common definition describes a positive question as a question where a positive response (“yes”) supports the truth of the hypothesis (Snyder & Swann, 1978; Klayman & Ha, 1987; Klayman, 1995). Posing positive questions does not however necessarily imply an ability to anticipate the epistemic effects of the “yes” or “no” responses. They can more simply originate from a tendency to inquire about features that “match” the hypothesis, i.e. features that are more typical of instances where the hypothesis is true, than of instances where it is false. When investigating whether a target individual is an extrovert, for example, asking “does she like parties?” is a positive question. The inquired feature matches the representation of an extrovert, and – as a result – a “yes” supports the hypothesis of extroversion, while a “no” weakens it. By contrast, asking “does she enjoy long solitary walks?” is a negative question: the feature matches the representation of an introvert, and accordingly a “yes” weakens the hypothesis of extroversion, and a “no” supports it. Symmetry/asymmetry of questions is more complex, as it is a matter of the quantity of information received and not only of its valence. An asymmetric query can – depending

on the answer it receives – confirm a hypothesis *more* than it can disconfirm it (asymmetrically *confirming* questions; or “high risk” testing strategies, Poletiek & Berndsen, 2000), or vice-versa (asymmetrically *disconfirming* questions; also known as “extreme” tests, Skov & Sherman, 1986; Slowiaczek, Klayman, Sherman, & Skov, 1992; or “low risk” testing strategies, Poletiek & Berndsen, 2000). Investigating, for instance, the extroversion of a person by asking “is she always the life of parties?” is an asymmetrically confirming test: a “yes” response is improbable, but, if received, would strongly support the hypothesis. On the other hand, a “no” response is probable, but – if received – only weakly disconfirms the hypothesis. Similarly the question “does she love spending most Saturday evenings reading poetry by herself?” is asymmetrically disconfirming (a “yes” strongly falsifies extroversion, whereas a “no” only weakly confirms it). For a stricter definition of asymmetry, many quantitative measures of the strength of confirmation are available (Crupi, Tentori, & Gonzalez, 2007). The first, most common, and easiest one is the Bayes’ factor. A common formal description of logically sound belief revision is the Bayes’ rule. It states how a degree of belief in the truth of a hypothesis H (expressed as the epistemic probability that H is true) should be revised in the light of a body of newly acquired evidence. A useful formulation of the Bayes’ rule is in terms of odds and likelihood ratios (Beyth-Marom & Fischhoff, 1983; Fischhoff & Beyth-Marom, 1983; Good, 1950, 1960, 1983; Osteen & Good, 1974):

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

where $p()$ is read “probability of”, H means “the hypothesis is true”, $\neg H$ means “the hypothesis is false”, E is the body of evidence, and the | symbol stands for a conditional probability (it can be read “given”). Reading from the left, the three terms of the formula are:

- (a) the posterior odds: the ratio between the probability that H is true given E and the probability that H is false given E;
- (b) the prior odds: the ratio between the probability that H was true before acquiring E, and the probability that it was false;

- (c) the Bayes factor, that is the Likelihood Ratio of E (thereafter, LR): the ratio between the probability of observing E assuming the truth of H and the probability of nevertheless observing E if H were false.

Bayes' rule is a straightforward and undisputed consequence of the basic axioms of standard probability calculus. More importantly, the LR appropriately describes an intuition that is common in many fields where correctly weighing evidence is critically important, such as medical diagnosis or legal judgement. A piece of evidence (e.g., a symptom, a clue) that is equally probable regardless of whether H (e.g., a possible diagnosis, a charge of wrongdoing) is true or false, does not change the probability that H is true or false, and therefore it is uninformative. Such a piece of evidence, with $p(E | H) = p(E | \neg H)$, has $LR = 1$, and thus leaves the posterior odds unchanged with respect to the prior odds. Along the same lines, a piece of evidence with $LR > 1$ increases the posterior probability of H with respect to $\neg H$: it is thus *confirmatory*. Finally, a piece of evidence with $LR < 1$ decreases the posterior probability of H with respect to $\neg H$: it is *disconfirmatory*. A dichotomous question – namely one accepting only “yes/no” as mutually exclusive answers – is symmetric if and only if the two answers have the same LR (a “yes” confirms H exactly as much as a “no” confirms $\neg H$, or vice-versa). Otherwise, it is asymmetric.

Symmetric queries are “fair” questions, with equal chances¹ of either confirming or disconfirming the hypothesis by the same amount of evidential strength. By choosing them, inquirers do not commit themselves either to a conservative or to a non-conservative stance (in technical terms, they equate the risk of incurring in a Type II, false negative, or a Type I, false positive, error). Asymmetric confirming tests have a relatively low probability of yielding strong evidence in support of the hypothesis, and a correspondingly high probability of finding weak evidence that refutes it. They are conservative questions, as they maximize the chances of (weakly) rejecting the hypothesis, while minimizing those of (strongly) accepting it. Asymmetric confirming

¹ Whilst the properties of LR's are independent of the priors $p(H)$ and $p(\neg H)$, the probabilities of receiving either a confirming or a disconfirming answer depend upon $p(H)$. Where we discuss them in this paper we assume that $p(H) = .5$, a common assumption in most other previous studies on this topic, and a premise in our experiments.

queries shift the balance in favour of Type II, false negative errors, and accordingly should be typical of contexts where there are good reasons to prefer type II errors to type I errors (e.g., when evaluating a crime charge in a judicial setting). By contrast, asymmetric disconfirming questions have a relatively low probability of finding strong evidence that disconfirms the hypothesis, and a correspondingly high probability of yielding weak evidence in support of it. By making probable a weak confirmation at the expense of an improbable strong refutation of the hypothesis, they denote a preference for risking Type I instead of Type II errors: a typical attitude of some preliminary medical screening tests (such as the PSA test for prostate cancer), or of the “overprotecting” policy in antiterrorism airport checks (Hammond, 2007). An alternative – and common – name for these latter sort of questions is “extreme tests” (Skov & Sherman, 1986), meaning that they address the feature that has the most extreme probability (either high or low) under the focal hypothesis, and the less extreme probability under the alternative one.

The properties of symmetric or asymmetric queries are formally independent of their positivity/negativity: that is, a question can be symmetric, asymmetric confirming, or asymmetric disconfirming, disregarding whether the response that supports the hypothesis is “yes” or “no”.

In theory, the best questions to posit are the most *diagnostic* ones, those with a maximal expected utility in informational terms. It can be measured as the mean LR – that is the weighed average of the LR of the confirming response in support of the hypothesis and the LR of the disconfirming response in support of the alternative hypothesis:²

$$MeanLR. = p(confirmation) * LR(confirmation) + p(disconfirmation) * \frac{1}{LR(disconfirmation)}$$

² Mean LR – or the mean Weight of Evidence (see Appendix C) – is a rough measure of the overall utility of a question, even though it is widely used in literature (e.g., Trope & Bassok, 1982; Slowiaczek, Klayman, Sherman, & Skov, 1992). A more proper measure is the less intuitive expected Information Gain (IG), computed as a difference in informational entropy before and after having received an answer to the question (Oaksford & Chater, 2007). In this case, as in most others where the prior probabilities of the hypotheses are the same, overall usefulness as measured by expected I.G. (or any other measure that has been proposed, such as Kullback-Leibler’s numbers) is directly proportional to the usefulness as measured by mean LR – that therefore is a viable measure.

By not choosing the most diagnostic questions among the available ones, a person risks to throw away useful information, thus increasing the chances of avoidable errors. Diagnosticity is never affected by the positivity/negativity of the query. Furthermore, it is not systematically affected by its symmetry/asymmetry: depending on the parameters associated to the tested features, there can be symmetric and asymmetric questions of equal diagnosticity, symmetric queries more diagnostic than asymmetric ones, or asymmetric queries more diagnostic than symmetric ones. The opportunity for deciding whether to address symmetric or asymmetric questions, and which type of the latter, on the grounds of context-driven preferences in order to risk different types of errors is granted. In order to minimize the *overall* probability of occurrence of an error of *any* type, however, these sorts of decisions should occur among questions of similar diagnosticity, wherever possible. Some psychologically-grounded preferences for certain testing strategies might foster the systematic selection of sub-optimal questions, thus unnecessarily increasing the overall chances of errors

These sorts of errors have been vastly studied in psychology under the name of “confirmation biases”. They might consist either of avoidable Type I errors (occurring when one accepts a false hypothesis as true on the ground of biased evidence), or of avoidable type II errors (when sticking to an old, improper, but believed true opinion, in the face of available contrary evidence; this is also known as “conservatism”, Fischhoff & Beyth-Marom, 1983; or “persistence of beliefs”, Nisbett & Ross, 1980), or of a mix of them. A wide debate has spawned on the formal and psychological details of how different testing strategies might result in biased acceptance or rejection of hypotheses. The issue is complex, even though it is very important for a better understanding of how different people sometimes get sincerely convinced of opposite and incompatible opinions, thus causing conflicts at all levels of society (Nickerson, 1998). Even though it can be argued that most of the observed psychological testing tendencies are not suboptimal *a-priori*, because they can be – and often are – “optimal” in some specific environments (e.g., Oaksford & Chater, 1994; McDonald, 1992), most scholars currently agree that some testing

strategies – including positive testing and asymmetric testing – can result in biased opinions mostly by their synergies with other psychological and environmental factors, such as different response sets (e.g. Zuckerman, Knee, Hodgins, & Miyake, 1995), information-evaluation biases (e.g. Lord, Ross & Lepper, 1979; Newman, Wolff, & Hearst, 1980; Ross & Anderson, 1982; Henrion & Fischhoff, 1986; Poses, Bekes, Copare & Scott, 1990; Slowiaczek, *et al.*, 1992; Koehler, 1993; Cameron & Trope, 2004), or the type of local environment where the evidence has been harvested (e.g. Klayman & Ha, 1987; Klayman, 1995; Nickerson, 1998; Fiedler, 2000; Fiedler, Brinkmann, & Betsch, 2000).

2. Empirical Evidence Concerning Asymmetry

There is a widespread acknowledgment that asymmetric questions are preferred to symmetric ones: “Subjects’ overall preferences in information gathering reflected an additive combination of three favored qualities: diagnosticity, positive testing, and extremity [i.e., asymmetrically disconfirming tests]” (Slowiaczek *et al.*, 1992, p.395); “three factors seem to drive people’s choices: diagnosticity, positivity, and extremity.” (McKenzie, 2004, p.206). Close scrutiny of the literature, nevertheless, shows that this consensus does not stand on firm ground. As far as the symmetric vs. asymmetric comparison is concerned, only a handful of data is available, and it is mostly inconclusive. Even the more specific comparison between preferences for asymmetric confirming vs. asymmetric disconfirming questions yielded partly contradictory findings.

2.1 Symmetric vs. Asymmetric Queries

Slowiaczek *et al.*’s (1992) wrote “there is evidence that people prefer to ask about features with one extreme probability, compared with symmetrical questions of equal diagnosticity”, citing as their only reference the “certainty bias” found by Baron *et al.* (1988). Baron and his colleagues, however, did not investigate symmetrical tests, apart from two very exceptional forms of them: certain tests, incapable of making errors, with $p(E|H)=1$ and $p(E|\neg H)=0$, and uninformative tests with $p(E|H)=p(E|\neg H)$. Results concerning those peculiar tests cannot be generalized to the totality

of symmetrical tests. Some more general data concerning preferences for symmetric vs. asymmetric tests can be extrapolated from studies that addressed different research issues, and they are mostly inconsistent with the view that asymmetric questions are generally preferred to symmetric ones. In their seminal study that proved the sensitivity of human hypothesis-testing strategies to the diagnosticity of questions, Trope and Bassok (1982) used, in their first Experiment, a set of four high-diagnosticity and four low-diagnosticity questions. Their research goal was to compare sensitivity to diagnosticity to the preference for positive testing. With reference to the probability parameters reported in Trope and Bassok's Table 1 (p.25), we classified their queries according to the symmetry/asymmetry dimension, as follows:

Low Diagnosticity questions:

1. very low $p(\text{feature}|\text{H})$: asymmetrically disconfirming, negative question
2. low $p(\text{feature}|\text{H})$: almost symmetric, negative question
3. high $p(\text{feature}|\text{H})$: almost symmetric, positive question
4. very high $p(\text{feature}|\text{H})$: asymmetrically disconfirming, positive question

High Diagnosticity questions:

5. very low $p(\text{feature}|\text{H})$: almost symmetric, negative question
6. low $p(\text{feature}|\text{H})$: asymmetrically confirming, negative question
7. high $p(\text{feature}|\text{H})$: asymmetrically confirming, positive question
8. very high $p(\text{feature}|\text{H})$: almost symmetric, positive question

Even though perfectly symmetric tests were not investigated, it is interesting that participants did not rate strongly asymmetric tests more than almost symmetric ones. In fact, the most preferred question (even though only slightly so) in each set was almost symmetrical (questions 3 and 5; Figure 1, p.26). Their second experiment used only perfectly symmetrical questions, and therefore it is not helpful for the present purpose. In their third experiment, features denoted by probability ranges instead of point probabilities were used. The most preferred and the least preferred questions were symmetrical (tagged respectively "high-high" and "low-low" in their Figure 4, p.30), with

another symmetrical question (tagged “intermediate-intermediate”) and all the asymmetrical ones laying in between. Overall, Trope and Bassok’s (1982) study does not suggest any preference for asymmetric testing. In later studies (Trope & Thompson, 1997; Cameron & Trope, 2004), it was demonstrated that when individuals try to make accurate judgements about an attitude (e.g. “opposing the killing of animals for their fur”) of a target person, the questions they generate are affected by their previous, stereotype-based expectancies. If the target person is a member of a social category that strongly implies the judged attitude (e.g., “vegetarians”), there is a preference for asymmetrically confirming questions (e.g., “do you oppose the killing of mice for research on life-saving pharmaceuticals?”). However, when the target person is non-stereotyped as far as the investigated attitude is concerned (e.g., “TV producers”), symmetrical questions were preferred (e.g., “do you hunt for sport?”).³ The authors argued that the increased *a-priori* expectancy that a stereotyped target shared the attitude raised the subjective probability of receiving the strongly diagnostic confirming answers to asymmetric questions. The increase in the probability of the strong answers proportionally increased the diagnosticity of asymmetric confirming tests. In conclusion, according to Trope and colleagues, the default tendency in absence of strong prior beliefs about the truth of the hypothesis is to prefer symmetrical, fair testing. Asymmetrical questions – of the confirming sort – are preferred only when a person is *a-priori* confident that the tested hypothesis is probably true.

In a very influential study, Skov and Sherman (1986) used “planetary” problems where participants were requested to test whether an extraterrestrial being belonged to one of two non-overlapped and *a-priori* equiprobable populations. In order to do so, participants had to select two yes/no queries about the alien’s features, whose probability distributions in the two populations were explicitly described. The authors built the questions by orthogonally crossing three factors:

³ Differently from other studies discussed in this section, Trope and Thompson’s (1997) and Cameron and Trope’s (2004) studies were not question-selection or question-evaluation tasks, but question-generation tasks. Subjects were free to generate any questions that they considered pertinent (apart from directly asking about the target attitude). Accordingly, explicit probabilistic parameters were not available *a-priori*, and symmetry/asymmetry of questions was evaluated *ex-post* by independent judges.

diagnosticity (low, medium, or high), positivity (positive vs negative), and “extremity” of a question, operationalized as the probability of receiving a hypothesis-confirming answer (high, medium, low). All of their questions with medium probability of receiving a confirming answer were perfectly symmetrical; those with a high probability of a confirming answer were asymmetrically disconfirming, and those with a low probability of receiving a confirming answer were asymmetrically confirming queries (Skov & Sherman, 1986, Table 2, p.106). Preferences focused upon asymmetrically disconfirming questions, followed by symmetric questions, and finally by asymmetrically confirming questions (respectively 40.0%, 34.3%, 25.8% of choices; data extrapolated from Skov & Sherman’s Table 4, p.110; the “both hypotheses” condition is not included in these figures). Even though the pattern shows a preference for asymmetric disconfirming questions over asymmetric confirming ones, a global preference for asymmetric tests over symmetric tests is apparently absent: symmetric queries, that were 1/3 of the available tests, accounted for 1/3 of the choices (34.3%). In Skov and Sherman’s “both hypotheses” condition, where the participants were told to test both the alternative hypotheses, thus not allowing to classify asymmetric questions into confirming or disconfirming ones, symmetric questions (again 1/3 of the total) accounted for 45.8% of the choices, against 54.2% of choices for asymmetric questions (2/3 of the total), showing – if anything – a marginal preference for symmetrical queries.

In Slowiaczek and colleagues’ (1992) study, Experiments 3a and 3b were the only ones investigating the selection of questions (Experiments 1 and 2 focused exclusively on the evaluation of responses). No symmetric questions were used in Experiment 3a (that used the “planetary” problems introduced by Skov and Sherman, 1986), even though some questions were less asymmetric than others (probability parameters are in Slowiaczek *et al.*, 1992, Table 1, p.393). The results replicated the preference for the most asymmetric disconfirming queries already shown by Skov and Sherman (1986), but are not useful for comparing preferences for symmetric and asymmetric questions, contrary to the authors’ claim. Conversely, in Experiment 3b some questions were almost symmetric, namely the “no extremity” queries in the contrasts 1, 2, 5, and 6 reported in

their Table 6, p.401. Those questions were compared to asymmetric questions, either asymmetrically disconfirming (comparisons 1 and 2) or asymmetrically confirming ones (comparisons 5 and 6). Furthermore, questions 1 and 5 were positive, and 2 and 6 were negative, thus allowing to compare the strength of positive testing to the preferences for symmetric/asymmetric testing. Unfortunately, results were ambiguous, and different between the two problem contents that were used (the planetary problems, and pseudo-medical problems). In comparisons 1 and 2, participants preferred asymmetrically disconfirming queries to symmetric queries, both in the positive as well as in the negative conditions, but only in the planetary context. In comparisons 5 and 6, negative asymmetrically confirming queries were preferred to negative symmetric questions in the medical context, but not in the planetary context. In both contexts, there were no differences between the asymmetric and symmetric positive questions (Slowiaczek *et al.*, 1992, Table 6, p.401).

To sum up, the actual empirical support for the alleged general preference for asymmetric over symmetric queries is scarce and contradictory. Some previous findings hint at a lack of preference either for symmetric or asymmetric questions, some others at a possible preference for symmetric questions in some contexts, and – eventually – others still emphasize a preference for asymmetric tests.

2.2 Asymmetric Confirming vs. Asymmetric Disconfirming Queries

As far as the two sorts of asymmetric queries are concerned, Trope and his colleagues' (1997, 2004) findings are contradictory with respect to those by Skov, Sherman, and their colleagues (1986, 1992). Whilst the former found that people prefer asymmetrically confirming queries when background knowledge suggests that there is a high a-priori probability that the tested hypothesis is indeed true, and symmetric queries otherwise, the latter found a general preference for "extreme tests" – namely, asymmetrically disconfirming queries (see 2.1). Poletiek and Berndsen (2000) made an attempt to reconcile those conflicting results by addressing hypotheses testing as a risk-taking behaviour, where the "risk" is that of disconfirming one's own hypothesis. The authors

argued that the choice of a testing strategy is mostly determined by motivational and contextual factors that suggest how to weigh the strength of the *confirming* evidence that is sought by the probability of actually finding it. Asymmetrically confirming queries are “high risk” tests, because they have a high probability of falsifying the hypothesis (however weakly). By contrast, asymmetrically disconfirming queries are “low risk” tests, because they have a high probability of confirming the hypothesis. Poletiek and Berndsen found that when testing scientific hypotheses and judicial hypotheses people prefer “high risk” strategies, consistently with the preference for asymmetrically confirming tests found by Trope and Thompson (1997) in inquiries concerning stereotyped targets.⁴ In those contexts people are apparently motivated to avoid type I errors as much as possible, at the cost of increasing the risk of type II errors, and choose their preferred tests accordingly. By contrast, in Skov and Sherman’s (1986; Slowiaczek *et al.*, 1992) “[...] unrealistic test situation, the subject may interpret the instructions as to find as many Fizos [i.e., the name of one of the two groups of extraterrestrial beings] as possible. The tester may thus be primarily concerned with the probability of finding some evidence rather than with the quality of the evidence. Consequently, low-risk tests (i.e. minimizing the risk of deciding that the creature is not a Fizo) are preferred to risky tests” (Poletiek & Berndsen, 2000, p. 111). In light of the available data – and of the possible role of background knowledge and motivational factors in shifting preferences – an excessive confidence in a “default” preference for asymmetric disconfirming queries seems premature.

3. Empirical Evidence Concerning Positivity

A default prevalence of positive testing vs negative testing is less controversial, but some conflicting results are still present in the literature. Preferences for positive tests were observed in the majority of empirical works that used Wason’s 2-4-6 task (1960) and selection task (1966,

⁴ A weakness in Poletiek and Berndsen’s study is that – instead of giving participants the explicit probabilistic parameters of different tests – they informally described the tests’ properties: participants were told to choose between seeking strong confirming evidence, with a low probability of actually finding it, or seeking weak confirming evidence, with a high probability of finding it. In our view, showing that people, when explicitly asked, say that they would rather look for strong evidence even though improbable, does not prove that they are able to tell a “high risk” test, when they actually see one.

1968). Most 2–4–6 studies showed a strong tendency to adopt a positive testing attitude. Only a subset of participants ever turned to negative testing, and most often they did so in later stages of the task, when ill-posed confidence had already been apportioned on an improper hypothesis (Kareev, Halberstadt, & Shafir, 1993; Klayman & Ha, 1987, 1989; Rossi, Caverni, & Girotto, 2001; Spellman, Lopez, & Smith, 1999; Vallée-Tourangeau, Austin, & Ramkin, 1995; Cherubini, Castelvechio, & Cherubini, 2005; Gale & Ball, 2006). We cannot fully account here for the huge amount of studies on Wason's selection task. However, apart from special settings where the antecedent and consequent of the tested rule are experienced as more common than their negations (Oaksford & Chater, 2003), or where the negated consequent is made more salient (e.g., Sperber, Cara, & Girotto, 1995), or in the various "deontic" versions of the task (e.g., Cheng & Holyoak, 1985), the negative testing strategy denoted by the explicit selection of the $\neg q$ cards is uncommon. For example, in an interesting version of the task where participants had to pay to acquire information and accuracy was incentivised, the $\neg q$ choices accounted for as few as 6.9% of all acquisitions, even though they were optimal selections for increasing the participants' own monetary outcomes (Jones & Sugden, 2001).

In question-selection and question-evaluation studies, however, findings concerning positive testing are more controversial. Trope and Bassok (1982) did not observe it in their Experiment 1, and only found a weak preference for positive tests in their Experiment 2. Baron *et al.* (1988) found a significant "congruence bias" – equivalent to positive testing – in their Experiments 1 and 2, where the participants focused on a given hypothesis; however, the bias weakened (Experiment 3) or disappeared (Experiments 4-5) where all the alternative hypotheses were made more salient to participants. Skov & Sherman (1986) found more compelling evidence of a preference for positive testing in question-selection tasks: 60.1% of selected questions in their study were positive (data extrapolated from their Table 4, p.110). In Slowiaczek *et al.*'s Experiment 3a a similar rate was found (63% Table 5, p.401). Unfortunately, neither Trope and Thompson (1997) nor Cameron and

Trope (2004) reported the proportion of questions generated by their participants that were positive or negative.

4. Goals of the Present Study

The present study is an empirical investigation of the asymmetric testing and positive testing strategies, aimed at clarifying some of the ambiguities that are present in previous literature. Because inconsistencies across the studies have been attributed to different methodologies and problem contents (e.g., Poletiek & Berndsen, 2000), we chose to use minimally contextualized, abstract question-selection or question-evaluation tasks, described by fully explicit probabilistic information (question-generation tasks are less precise as far as the control of probabilistic and informational parameters are concerned). In light of the surprising lack of direct empirical evidence in support of the widespread idea that asymmetric testing is generally preferred to symmetric testing, our first goal was to directly compare preferences for symmetric and asymmetric questions (Experiments 1 to 3). By orthogonally factoring the positivity/negativity of questions, we also compared the relative strengths of the positive testing and asymmetric testing tendencies, and gathered further data concerning the extent of the positive testing tendency in question-selection (Experiment 1) and question-evaluation (Experiments 2, 3, 4) tasks. Finally, we purported to check whether preferences among asymmetric questions privileged either the confirming (e.g., Trope & Thompson, 1997) or the disconfirming ones (e.g., Skov & Sherman, 1986) (Experiments 1 to 4). We strived to control most of the formal properties of the questions that might have had a bearing on their choice, as much as possible. Scholars interested in the replication or reinterpretation of the results will find examples of the instructions and all the relevant parameters of the different problems that we used in extensive Appendices (A to G).

5. Experiments 1 and 2

5.1 Materials, Design and Procedure

Experiments 1 and 2 differed only in their dependent variable and – consequently – in part of the instructions. The procedure, materials, and experimental design were the same, and the

participants were drawn from the same pool. In each of a series of eight paper-and-pencil problems two decks of cards were described, “Deck A” and “Deck B”. Each one had 100 cards. Each card showed zero to four independent geometric figures, selected among triangles, circles, squares or pentagons. In each problem a table synthesized the distribution of cards with each figure in each deck (see an example in Appendix A).

Participants were told that the experimenter drew a card at random from a deck. They were then shown four questions, concerning the presence of the four features on the card: “is there a triangle on the card?”, “is there a square on the card?”, “is there a circle on the card?”, “is there a pentagon on the card?”. They had to select (Experiment 1) or rank in order (Experiment 2) the questions that they deemed most useful for surmising whether the card was more probably drawn from Deck A (for one group of participants), or from Deck B (for a second group). In all the problems the distribution of figures in the decks was such that two questions were symmetric, and two asymmetric. Orthogonally, in each problem two questions were negative, and two were positive. Accordingly, in each problem there was a symmetric positive question, a symmetric negative question (with LR values equal to the previous one), an asymmetric positive question, and an asymmetric negative question (with LR values equal to the previous one). The eight problem versions (see Appendix B) originated from balancing whether the asymmetric questions had either a larger, smaller or equal LR (and Information Gain, IG) than the symmetric questions, for each possible answer (same LR for “yes”, decreased LR for “no”; same LR for “yes”, increased LR for “no”; same LR for “no”, decreased LR for “yes”; same LR for “no”, increased LR for “yes”; LR decreased for both responses; LR increased for both responses; LR increased for “yes”, and decreased for “no”; LR decreased for “yes”, and increased for “no”; see the eight problems in the Appendix C). Thanks to this systematic balancing, observed preferences either for symmetric or asymmetric questions – averaged across the eight problems – could not be caused by a perceived increment or decrement of the actual strength of the possible answers to questions. As a further caution, we balanced (even though not orthogonally with respect to the previous balancing), the

association of symmetric and asymmetric questions with extreme probabilities: that is, in four problems the most extreme probabilities (considering both hypotheses) were associated with symmetric questions, whereas in the other four problems they were associated with the asymmetric ones. This manipulation also obtained an orthogonal balancing, across the problems, of the positivity/negativity of an asymmetric test and its being either confirmatory or disconfirmatory (four questions each for positive confirming, positive disconfirming, negative confirming, and negative disconfirming queries; see Appendix C). Unfortunately, having given priority to these balancing factors for theoretical reasons (i.e., according to Trope and colleagues, 1997, 2004, people can be sensitive to increments or decrements in the informative value of an answer; and, according to Skov & Sherman, 1986, and Slowiaczek and colleagues, 1992, people are mostly insensitive to the LR of answers, but are attracted by questions with extreme probabilities), we could not fully balance the diagnosticity of the questions. In five problems, symmetric questions were on average more informative than asymmetric questions, whereas in only three problems asymmetric questions were more informative than symmetric ones (see Appendix C). However, we planned to statistically check *post-hoc* whether this formal parameter affected choices.

Booklets containing general instructions (comprising two easy examples) and the eight experimental problems (one per page) were handed out individually to each participant. The pseudo-random order of presentation of the eight problems was different for each participant. As a further balancing factor, in each Experiment half of the participants received problem versions where they were asked to check whether the card had been drawn more probably from Deck A than from Deck B (focal hypothesis: Deck A), whereas the other half was asked to check whether the card had been drawn more probably from Deck B than from Deck A (focal hypothesis: Deck B). Thanks to this manipulation, positive and negative questions were matched for diagnosticity between groups: each question that was positive when focusing on Deck A was negative when focusing on Deck B, with most of its remaining characteristics (symmetry, extremity, diagnosticity) being kept constant (apart from being confirmatory vs disconfirmatory: e.g., asymmetric positive

confirming questions for Deck A were asymmetric negative disconfirming ones when focusing on Deck B, and so on; see appendix C).

Participants were asked individually for informed consent, and, if they consented, were tested individually or in small groups (in the latter case, they could not consult each other) in quiet environments (university libraries or study rooms). They were told that they could proceed at their own pace in responding to the problems.

5.2 Experiment 1

5.2.1 Participants.

A total of 30 volunteers (15 female, 15 male, mean age: 23.8 years, range 18-28 years; mean education: 17, SD 1.9) took part in the study. They were mostly undergraduate students from the university of Milano-Bicocca.

5.2.2 Dependent variables.

Participants were instructed to select, out of the four that were presented, the two questions that they deemed most useful. This was not a ranked, but an “all or none” judgement, similar to the dependent variable used by Skov and Sherman (1986), but different from that used by Trope and Bassok (1982). Secondly, for each problem participants were asked to express their degree of confidence in the correctness of their selections on a rating scale graded 1 to 7, with 1 corresponding to “least confident” and 7 to “most confident”.

5.2.3 Results and analyses.

Table 1 shows the raw number of total choices (and proportions) as a function of positivity and asymmetry of the questions.

insert Table 1 about here

There is no reliable preference for symmetric questions over asymmetric ones, or vice-versa (exact binomial test; symmetric questions: 239; asymmetric questions: 241; $p = 0.96$). By contrast,

positive questions were preferred to negative ones (positive questions: 375; negative questions: 105; $p < 0.001$). The two factors did not interact ($\chi^2 = 0.08$; $df = 1$; exact $p = 0.83$). More appropriate analyses, however, considered the six possible patterns of choices available to participants for each problem: choosing the two positive questions, the two negative ones, the two symmetric ones, the two asymmetric ones, the positive symmetric one plus the negative asymmetric one, or the positive asymmetric one and the negative symmetric one. Each participant was assigned a score ranging from 0 to 8 for each pattern of choice, by counting its occurrence over the eight problems. Means and standard deviations are shown in table 2.

insert Table 2 about here

Table 3 shows the exact p values of each comparison among the six possible patterns, by means of Wilcoxon non-parametric exact tests.

insert Table 3 about here

Results show three ranks of preferences. Selecting the two positive questions (one symmetric and one asymmetric) in each problem is the most common tendency, followed by selecting either the two symmetric or the two asymmetric questions (one positive, and one negative), with no reliable differences between the two of them. Other choices, consisting in choosing either the two negative questions, or the positive symmetric question and negative asymmetric question, or the positive asymmetric question and negative symmetric question, are residuals, without differences among the three of them.

Table 4 shows the raw numbers (and proportions) of *asymmetric* choices as a function of positivity/negativity and of the type of asymmetry (confirmatory vs disconfirmatory).

insert Table 4 about here

There is no reliable preference for either asymmetric confirming or asymmetric disconfirming questions (120 vs 121 choices). Positive choices are reliably more than negative ones, consistently with overall results (exact binomial test, $p < 0.001$). The interaction is not significant ($\chi^2 = 2.28$; $df = 1$; exact 2-tailed $p = 0.16$).

Correlations with measures of diagnosticity were computed on data aggregated across participants. The number of times each question was chosen correlated reliably with both measures of diagnosticity that we used: mean Weight of Evidence (Spearman's $r = 0.53$, $N=32$, $p<0.005$) and expected Information Gain ($r = 0.54$, $p<0.005$).

Mean confidence across the problems did not correlate reliably with the score attained in each one of the six possible patterns of choice. However, mean confidence correlated positively (Pearson's $r = 0.47$, $p < 0.01$) with the number of symmetric positive questions selected by each participant, and negatively ($r = -0.40$, $p < 0.05$) with the number of negative asymmetric questions selected by each participant.

5.2.4 Discussion.

The results are clear cut, and need little interpretation. Most notably, we observed a strong tendency to select positive questions. The more positive symmetric questions people choose, the more they were confident in the correctness of their choice, and vice-versa for negative asymmetric questions. Secondly, there was a weaker tendency to select either two symmetric questions, or two asymmetric ones, with no differences between them. Among the asymmetric choices, there is no apparent preference either for confirming or for disconfirming questions. Overall, results confirm and strengthen previous literature concerning the preference for positive over negative questions in hypothesis testing. The correlation between the diagnosticity of each question in each problem and the number of times it was chosen – consistent with previous literature (e.g., Trope & Bassok,

1982) – does not invalidate the finding concerning preferences for positive tests, because positive and negative queries were balanced for diagnosticity between groups. On the other hand, sensitivity to diagnosticity might have shifted preferences toward symmetric questions. In five problems out of eight, symmetric questions were more informative than asymmetric ones. This notwithstanding, no significant preference either for symmetric or asymmetric questions was observed. Furthermore, no differences between the asymmetric confirming and disconfirming questions were observed. These negative findings are in contrast with those discussed in the introduction: Trope and his associates (Trope & Thompson, 1997; Cameron & Trope, 2004) found preferences for symmetric questions where there was no strong *a-priori* belief in the hypothesis, and preferences for asymmetrically confirming questions where there was a strong *a-priori* belief in the hypothesis; Poletiek and Berndsen (2000) found preferences for asymmetrically confirming questions in the scientific and judicial domains; Skov and Sherman (1986) found preferences for asymmetrically disconfirming questions in problems contextualized as planetary explorations; Slowiaczek *et al.* (1992) found the same in the same context, but ambiguous results when comparing the planetary context to pseudo-medical problems. These findings, being negative, will need further replication before any discussion of the inconsistencies is warranted.

5.3 Experiment 2

The dependent variable in Experiment 1 – choosing the two best questions – together with the observed strong preference for choosing the two positive questions (one of which was symmetric, and the other asymmetric), might have concealed some subtle differences of preference for symmetric and asymmetric questions. Experiment 2 explores this possibility. It is identical to Experiment 1 except for its dependent variable: instead of asking participants to choose the two best questions, we asked them to rank the four questions in order of usefulness (1 = most useful, 4 = least useful), a procedure that is similar to that used by Trope and Bassok (1982). In the instructions for the task we clarified that ties could be indicated by assigning the same rank to two or more questions.

5.3.1 Participants.

A total of 30 volunteers (16 female, 14 male, mean age: 23.8 years, range 19-38 years; mean education: 16, SD 2.5) took part in the study.

5.3.2 Results and analyses.

Table 5 shows the mean ranks assigned by participants to each question (averaged over the eight problems). A set of two-tailed Wilcoxon exact tests confirmed that positive questions were preferred to negative questions ($p = 0.001$), but it also showed that symmetric questions were apparently preferred to asymmetric ones ($p = 0.002$).

insert Table 5 about here

Positive symmetric questions were rated more useful than positive asymmetric ($p = 0.002$), negative symmetric ($p < 0.001$) and negative asymmetric ones ($p < 0.001$). Positive asymmetric questions were preferred to negative symmetric ones ($p = 0.03$) and negative asymmetric ones ($p = 0.002$). Finally, negative symmetric questions were ranked more important than negative asymmetric ones ($p = 0.04$). By aggregating data across participants, however, we found reliable negative correlations between the mean rank attributed to each questions and its diagnosticity (Pearson's $r = -0.58$, $N=32$, $p < 0.001$; results do not differ among the two alternative measures of diagnosticity that we used). This finding, together with the fact that, in the stimuli, symmetric questions were more diagnostic than asymmetric ones in five problems out of eight, might have led to an artificial conflation of the observed preference for symmetric questions. A correlation might of course also be read the opposite way around: preferences for symmetric questions, that had a slight advantage in diagnosticity, could have contributed to the observed correlation. In order to disentangle these two interpretations we divided post-hoc the 8 problems in two groups: those where symmetric questions were more diagnostic (problems 2 to 6), and those where they were less

diagnostic (problems 1, 7, 8). The mean ratings of symmetric and asymmetric questions in the two groups are reported in Table 6.

insert Table 6 about here

Two two-tailed Wilcoxon exact tests showed that symmetric questions were significantly preferred in the problems where they were more informative ($p < 0.001$), but asymmetric ones were preferred in the problems where the symmetric questions were less informative ($p = 0.003$). These post hoc tests are spurious, since dividing post-hoc the problems in two unequal groups sacrifices some of the balancing factors. Considering however that their results are consistent with the previously known tendency to prefer questions with a high diagnosticity (e.g., Trope & Bassok, 1982; Garcia-Marques, Sherman, Palma-Oliveira, 2001), in our view it is acceptable evidence that the apparent preference for symmetric questions reported above is – at least partly – an artifact generated by the differences in the diagnosticity of the problems.

Table 7 shows the mean ranks assigned by participants to asymmetric questions only, as a function of their positivity/negativity and of their type (confirmatory vs disconfirmatory).

insert Table 7 about here

Two two-tailed Wilcoxon exact tests confirmed, once again, that positive questions were preferred to negative ones ($p < 0.001$), whereas there were no significant differences between confirming and disconfirming asymmetric queries.

Finally, there were no significant correlations among mean confidence and the ranks assigned to each question, or their diagnosticity.

5.3.3 Discussion.

The results of Experiment 2 mostly replicate those of Experiment 1, thus weakening the concern that the lack of preferences either for symmetric or asymmetric questions might have been induced by the type of dependent variable used in the former experiment. The preference for positive questions is replicated, and so is the lack of preference for asymmetric questions. Even though raw data show a preference for symmetric questions, this finding is probably an artefact accounted for by the participants' sensitivity to the diagnosticity of the questions. Among asymmetric questions, there are no preferences either for confirmatory or disconfirmatory ones, in keeping with Experiment 1 and contrary to previous studies showing such preferences (see the discussion of Experiment 1 and the introduction). Apparently, in tasks where the effects of domain-related previous knowledge is minimized, people are indifferent to whether a question is symmetric or asymmetric, and all the more they are indifferent to whether an asymmetric question is confirmatory or disconfirmatory. By contrast, even in those tasks people are sensitive to the diagnosticity of questions, and they are appealed by positive questions.

6. Experiment 3

Klaus Fiedler, commenting on a previous draft of this study, correctly noticed that, with the exception of problem 3, the numerals describing the asymmetric queries in Experiment 1 and 2 were more complex than those describing the symmetric queries, that were rounded to the tens (e.g., 80%, 20%, 90%, vs 95%, 54%, 46%, etc.; see Appendix B). Rounded numerals might have been more appealing to participants, and easier to process. This feature might have contributed to the shift of preferences to symmetric questions, and might have concealed possible preferences for asymmetric testing. In Experiment 3 we removed that confounding. The Experiment's design and procedure were exactly the same as in Experiment 2, but we used new problems, none of which used rounded numerals in their probability parameters. The new problems and their parameters are reported in Appendices D and E.

6.1 Participants

A total of 48 volunteers (29 females, 19 males, mean age: 23.4 years, range 20-34 years; mean education: 16.5 years, SD 1.6) took part in the study. They were mostly undergraduate and graduate students from the University of Milano-Bicocca (Northern Italy) and from the University of Chieti (Southern Italy).

6.2 Results and Analyses

Table 8 shows the mean ranks assigned by participants to each question (averaged over the eight problems).

insert Table 8 about here

A set of two-tailed Wilcoxon exact tests confirmed that positive questions were preferred to negative questions ($p < 0.001$), and that symmetric questions were apparently preferred to asymmetric ones ($p = 0.03$), fully replicating the main findings of Experiment 2. Positive symmetric questions were rated marginally more useful than positive asymmetric ones ($p = 0.06$), and significantly more useful than negative symmetric and asymmetric ones ($p < 0.001$ for both comparisons). Positive asymmetric questions were preferred to negative symmetric and asymmetric ones ($p < 0.001$). There were no differences between negative symmetric and asymmetric questions. As occurred in the previous two experiments, however, the diagnosticity of questions possibly contributed to choices, as shown by a significant negative correlation between the mean rating of each question (across participants) and its diagnosticity ($r = -0.55$, $p < 0.001$; the results are the same for the two alternative measures of diagnosticity that we used). Accordingly, we compared symmetric and asymmetric queries in the two subgroups of problems where symmetric queries were more informative (problems 2 to 6) or less informative (problems 1, 7, 8). Similarly to Experiment 2, we found that each type of question was reliably preferred in the group of problems where it was more diagnostic (Wilcoxon exact tests; $p = 0.003$ for the five problems where symmetric questions

are more informative, and $p < 0.001$ where asymmetric questions are the more informative ones).

Mean ratings are shown in Table 9.

insert Table 9 about here

With the caveats associated to this way of splitting problems post-hoc into two unbalanced groups, this finding suggests that the preference for symmetric questions might – at least partly – be an artefact induced by the different diagnosticity of queries.

Table 10 shows the mean ranks assigned by participants to asymmetric questions only, as a function of their positivity/negativity and of their type (confirmatory vs disconfirmatory).

insert Table 10 about here

Two two-tailed Wilcoxon exact tests confirmed that positive questions were preferred to negative ones ($p < 0.001$), whereas there were no significant differences between confirming and disconfirming asymmetric queries, in full compliance with the previous Experiments.

There was no significant correlation among mean confidence and the rank assigned to each question.

6.3 Discussion

In Experiment 1 and 2 symmetric questions were described by easy probability parameters rounded to the tens, whereas asymmetric questions were almost always described by more complex numerals. This might have induced a shift of preference towards symmetric questions, that in turn might have concealed preferences for asymmetric testing. In Experiment 3 we removed that concern, and nonetheless we found no preference whatsoever for asymmetric tests. If anything, symmetric tests were again preferred, however – as occurred in Experiment 2 – this could have been a collateral effect of their slightly greater diagnosticity, that significantly correlated with the

participants' answers. Focusing on asymmetric queries, once again we did not find any reliable difference of preferences between confirmatory and disconfirmatory ones.

7. Experiment 4

In the three previous Experiments, two negative findings were recurrently observed, suggesting that they are worth of notice. Firstly, the lack of preferences for asymmetric queries over symmetric ones conflicts with previous claims (e.g., Slowiaczek *et al.*, 1992). In fact, however, it is not in strong contrast with empirical data of previous studies that were inconclusive as far as this comparison was concerned (see the introduction). The second negative result is more puzzling. Previous studies consistently showed that – where asymmetric questions were chosen or generated – systematic preferences were observed either for confirmatory (Trope & Thompson, 1997; Cameron & Trope, 2004) or for disconfirmatory queries (Skov & Sherman, 1986; Slowiaczek *et al.*, 1992). This was not the case in the Experiments 1, 2 and 3. In aggregated form, participants' choices were indifferent to the confirming or disconfirming valence of the asymmetric questions. The designs of Experiments 1 to 3 might not have afforded the necessary power for detecting subtle differences between different sorts of asymmetric queries. The presence of two symmetric and two asymmetric questions in each problem implies that participants who ranked the two positive questions as the first and second in order of importance (a strong tendency, as shown by the comparisons between positive and negative queries) were also ranking highest a symmetric *and* an asymmetric question, the latter being, across problems and groups of participants, half the time confirming, and half the time disconfirming. This feature might have led to an artifactual balance of the preferences for confirming and disconfirming asymmetric questions. In order to remove this possible confusion, in Experiment 4 we did not use symmetric questions. All of the four questions available in each problem were asymmetric, two of them negative and two positive, and – across problems – half of them confirmatory, and the other half disconfirmatory. The goal was to detect preferences either for confirming or disconfirming questions that were possibly concealed by previous designs.

7.1 Participants

A total of 48 volunteers (29 females, 19 males, mean age: 23.4 years, range 20-34 years; mean education: 16.5 years, SD 1.6) took part in the study. They were mostly undergraduate and graduate students from the University of Milano-Bicocca and from the University of Chieti.

7.2 Design and Procedure

The procedure, setting and instructions to the participants were the same as in Experiments 2 and 3. Stimuli were built by systematically varying the value of the most extreme probability of each feature under the two alternative hypotheses, and its association either with the focal hypothesis, or with the alternative one. For the former manipulation, probabilities were classified into four groups: very high ($p > 0.9$), high ($0.5 < p < 0.9$), low ($0.5 > p > 0.1$), and very low ($p < 0.1$). A feature probability is extreme when it is further removed from .5 than the corresponding probability under the alternative hypothesis. For example, if a circle has a probability of 0.68 of being present in Deck A, and 0.18 in Deck B, and the participant is focusing on Deck A, the most extreme probability (0.18) is in the “low” range, and is associated to the non-focal hypothesis. As in another example, if $p(\text{square}|\text{Deck A}) = 0.98$, and $p(\text{square}|\text{Deck B}) = 0.89$, the extreme probability (0.98) is in the “very high” range, and is associated to the focal hypothesis (see Appendix F). Questions classified as “low” or “very low” ranges of extreme probability are negative and disconfirming if the extreme is associated to the focal hypothesis. They are positive and confirming if the extreme is associated to the non-focal hypothesis. Vice-versa, questions classified as “high” or “very high” are positive and disconfirming if the extreme is associated to the focal hypothesis and negative and confirming if it is associated to the non-focal hypothesis. Thus, by manipulating the association of the extreme probability either to the focal or to the non-focal hypothesis, and its magnitude, we also attained the basic positivity/negativity x type of asymmetry design (see Appendix G). Furthermore, by varying probabilities over four magnitudes (instead of two: i.e., $p > 0.5$ vs $p < 0.5$), we purported to explore whether the absolute probability values also had an effect on preferences. In order to fully balance the questions in the four ranges of extreme probability we had to devise two sets of

four problems each, that – unavoidably – also differed between them in diagnosticity. Four problems had highly diagnostic questions, whilst the other four problems had less diagnostic questions (expected W.E.: 4.15 vs 0.87). However, the diagnosticity of questions within each problem was kept constant.

7.3 Results and Analyses

Results are shown in Table 11.

 insert Table 11 about here

Disconfirmatory questions were preferred to confirmatory ones, as found by Skov and Sherman (1986) and Slowiaczek *et al.* (1992), only for positive questions. For negative questions, confirmatory tests were deemed more important than disconfirmatory ones. The positivity by type of asymmetry interaction that is apparent in the table is best described by the following pattern of significant differences: positive disconfirming questions $>_{p=0.011}$ positive confirming $>_{p=0.014}$ negative confirming $>_{p=0.034}$ negative disconfirming (all of the remaining pairwise comparisons between the four types of questions were also significant; these and the following p values – unless otherwise stated – were obtained from 2-tailed Wilcoxon exact tests). Quite against immediate intuition, this pattern indeed suggests that participants were little – if at all – affected by the confirming or disconfirming valence of an asymmetrical question. They were more likely affected by the magnitude of the probability of a feature under the focal hypothesis $p(E|H)$. Elementary algebra shows that, diagnosticity of questions being equal, positive disconfirming questions have necessarily a $p(E|H)$ greater than positive confirming questions, and negative confirming questions have necessarily a $p(E|H)$ greater than negative disconfirming ones.⁵ Namely, the most preferred positive disconfirming questions had, on average, a $p(E|H)$ greater than the second-rated positive

⁵ In our design, some positive confirming questions had a $p(E|H)$ greater than some positive disconfirming questions occurring in different problems, but this occurred because those questions had different diagnosticity (see Appendices F and G).

confirming ones. The third rank in order of importance was assigned to negative confirming questions, with a $p(E|H)$ greater than that of negative disconfirming ones. The latter were deemed the least important queries. Analyses considering the four ranges of magnitude of the extreme probabilities and their association either to the focal or to the alternative hypothesis corroborated and further specified this reading. The results are shown in Table 12.

insert Table 12 about here

In the very high and high extremity ranges, questions were deemed more important where the extreme was associated to the focal hypothesis (positive disconfirming questions), than to the non-focal one (negative confirming questions) ($p = 0.001$ and $p < 0.001$, respectively). Very high-extreme and high-extreme questions did not differ between them when they were positive, whilst there was a preference for very high-extreme over high-extreme negative questions ($p = 0.044$). In the low and very low extreme-probability ranges, the preferred questions were those whose extreme probabilities were associated to the non-focal hypothesis (positive confirmatory questions) ($p < 0.001$ and $p = 0.005$, respectively). The very low-extreme and low-extreme questions did not differ between them when they were associated to the focal hypothesis (i.e., they were negative), whilst they differed significantly when they were associated to the non-focal hypothesis (and hence, they were positive questions, the low extreme questions with $p(E|H)$ higher than very low extreme questions; $p = 0.003$). Altogether, it seems that it was not extremity under the focal hypothesis that drove participants' preferences, *pace* Skov & Sherman (1986): more simply, participants paid heed to the magnitude of $p(E|H)$, considering mostly unimportant those questions where $p(E|H)$ was very low, and increasingly important questions with higher $p(E|H)$, both in the negative and in the positive domain. In line with the previous considerations, $p(E|H)$ significantly correlated with the mean rank assigned to each one of the 32 questions (Pearson $r = -0.71$, $p < 0.001$), whereas $p(E|\neg H)$ did not ($r = -0.18$, not significant), also implying that participants were not maximizing $p(E)$ in

general, but specifically focused on $p(E|H)$. Even though unorthodox for ordinal data, linear regression is helpful for describing the two main findings of Experiment 4. A linear regression model that uses as predictors $p(E|H)$ and the positivity of the question [operationalized into a continuous measure by computing the difference $p(E|H) - p(E|\neg H)$, obtaining positive values for positive questions and negative values for negative ones] significantly fits the data ($r = 0.77$; $r^2 = 0.59$; $p < 0.001$), with the following parameters:

$$\text{mean rank} = 2.28 - 0.57 p(E|H) - 0.42 [p(E|H) - p(E|\neg H)]$$

By adding to the model, as predictors, $p(E|\neg H)$, the LR of the confirming answer, the LR of the disconfirming answer, or any combination of them, fitness is not significantly improved.

Finally, analyses of the confidence ratings did not yield any interesting result.

7.4 Discussion

Experiment 4 confirms the preference for positive questions already observed in the previous Experiments. It also shows that, for asymmetric questions, positivity/negativity interacts with the confirming/disconfirming valence of the question. Positive disconfirmatory questions are rated more important than positive confirmatory ones, whereas in the negative domain the opposite is true, with confirmatory questions ranking above disconfirmatory ones. This pattern of preferences seems inconsistent with the alleged preference to test features with extreme probabilities under the focal hypothesis, as suggested by Skov and Sherman (1986). Extremity under the focal hypothesis attracts preferences only when the extreme value is high or very high. When it is low or very low, questions are preferred where the extreme probability is associated to the non-focal hypothesis, thus resulting in positive, confirming questions. In our view, the pattern is best described by assuming that participants' preferences were mostly driven by two easily accessible heuristic clues: positivity of the question, that is $p(E|H) > p(E|\neg H)$, and the magnitude of $p(E|H)$, as described by a linear regression where these two features were the best predictors of responses. In these abstract tasks participants' choices – on the whole – are not strongly affected by sophisticated features such as the confirming or disconfirming valence of a question (that would

suggest that they are estimating *ex-ante* the epistemic strengths of its answers), or by the extremity of a feature's probability under one of the two hypotheses (that would suggest that they are trying to maximize either the probability of a confirming, or of a disconfirming response). More simply, participants prefer testing features that are *typical* of true instances of the focal hypothesis: typical both in a relative sense, as shown by preferences for questions where $p(E|H) > p(E|\neg H)$ (i.e., positive testing), as well as in an absolute sense, as shown by the increasing preferences for questions with increasing $p(E|H)$.

8. General Discussion

The study of spontaneous hypothesis-testing strategies has a long history in cognitive psychology, dating back to its early years (e.g. Wason, 1960). It can improve our understanding of how beliefs and opinions are – properly or improperly – built, maintained, or rejected. Preferences for different sorts of questions, by interacting with the environment and with psychological tendencies affecting the generation and evaluation of responses, sometimes result in loss of information and in an increased probability of errors, namely “confirmation biases” (e.g., Nickerson, 1998; McKenzie, 2004). Currently we know a great deal about human hypothesis-testing strategies, and about their consequences. Yet, some issues still remain obscure. First of all, close scrutiny of previous studies showed that they are inconclusive with regards to the alleged (e.g., Slowiaczek *et al.*, 1992) preference for asymmetric questions – i.e., questions that can confirm a focal hypothesis more strongly than they can falsify it, or vice-versa – over symmetric ones. Secondly, studies on preferences for different sorts of asymmetric questions have conflicting results: some reported preferences for confirmatory queries (e.g. Trope & Thompson, 1997), as opposed to others reporting preferences for disconfirmatory ones (e.g., Skov & Sherman, 1986). Poletiek and Berndsen (2000) attributed these differences to context-driven motivations and prior knowledge. As a matter of fact, across studies, the problem contents ranged from social inferences concerning the attitudes of a target person (e.g., Trope & Thompson, 1997), to planetary explorations (e.g., Skov & Sherman, 1986), to pseudo-medical, pseudo-scientific, or pseudo-judicial

problems (e.g., Slowiaczek *et al.*, 1992; Baron *et al.*, 1988; Poletiek & Berndsen, 2000), possibly raising a host of different prior knowledge-based considerations. Thirdly and finally, even though evidence in support of positive testing – i.e., a preference for testing features that are more consistent with the truth of a focal hypothesis than with its falsity – is less controversial (but see Trope & Bassok, 1982, and Baron *et al.*, 1988), its relative strength has not been compared to asymmetric testing yet. Such a comparison is theoretically relevant. Asymmetric testing strategies are attributed either to the ability to anticipate the confirming or falsificatory strength of an answer, and to balance it with its subjective probability (Trope & Thompson, 1997; Cameron & Trope, 2004; Poletiek & Berndsen, 2000), or to the ability to estimate and maximize the probability of receiving a confirming answer (Skov & Sherman, 1986; Slowiaczek *et al.*, 1992). They require fairly sophisticated cognitive judgements. By contrast, positive testing is a quite simple strategy: it checks for features that are more typical of the focal hypothesis than of its alternative. Comparisons of the relative strengths of positive and asymmetric testing can accordingly help understand whether spontaneous hypothesis-testing strategies in unfamiliar, abstract settings, are mostly intuitive judgments, or mostly analytical ones, in terms of the contraposition proposed by the current dual-process theories of human thinking (e.g., Gilovich, Griffin, & Kahneman, 2002).

In four experiments we addressed the three issues described above. We minimized the role of previous knowledge and contextual factors by using unfamiliar, abstract tasks. Participants had to select or evaluate questions about a card's features, in order to infer from which of two decks it had been drawn. The probability of occurrence of each feature in the cards of each deck was explicitly given. In Experiments 1 to 3 we systematically compared positive symmetric, positive asymmetric, negative symmetric, and negative asymmetric questions, while balancing the informational value of the responses, the confirming or disconfirming nature of the asymmetric queries, and the association of very high or very low probabilities either to symmetric or to asymmetric queries. In Experiment 1 participants had to select the two most important questions in each problem (similarly to Skov & Sherman, 1986), whereas in Experiment 2 and 3 they had to rank the questions in order

of importance (similarly to Trope & Bassok, 1982). Experiment 3 adjusted for a possible confound present in Experiments 1 and 2, by balancing the complexity of the numerals that described the probabilities associated to each feature. The results were consistent throughout the three experiments. A strong preference for positive questions was observed. By contrast, no preference for asymmetric questions was observed. If anything, there were significant preferences for symmetric questions (Experiments 2 and 3), even though they were probably an artefact originating from a slight advantage in diagnosticity of symmetric over asymmetric queries. By focusing analyses on asymmetric queries only, we did not find any systematic preference either for confirming, or for disconfirming queries. In Experiments 1 to 3, aggregated data showed that people were sensitive to the positivity/negativity of questions, and to their diagnosticity, but that they were mostly unaffected by the symmetry/asymmetry, and – internally to asymmetric questions – by their disconfirming or confirming valence. However, Experiments 1-3, specifically designed for comparing positive testing to asymmetric testing, might not have afforded enough power to detect subtle differences between asymmetric queries of different sorts. Experiment 4 complemented them by investigating asymmetric queries only, and by varying systematically their positivity/negativity, their confirmatory/falsificatory valence, and, embedded in that design, the magnitude of the probability parameters describing each questions. Its results confirmed a robust preference for positive questions over negative ones. They also showed an interaction: positive disconfirming questions were preferred to positive confirming ones, whereas negative confirming questions were rated more important than negative disconfirming ones. This trend – as far as we know – was never observed before: previous studies either detected a preference for confirmatory questions (e.g., Trope & Thompson, 1997), or for disconfirmatory questions (e.g., Skov & Sherman, 1986), but we found no reports of a preference reversals depending on the positivity/negativity of the question. Close scrutiny of the data suggested that this interaction originated from the participants' preference for testing features with relatively high probabilities under the focal hypothesis, a heuristic behaviour reminiscent of “pseudodiagnostic” judgements (e.g., Doherty, Mynatt, Tweney, &

Schiavo, 1979; Doherty & Mynatt, 1986), and of the “sufficiency” strategy in the evaluation of contingency tables (Mandel & Lehman, 1998). Both those phenomena are denoted by a focus of participants on conditional probabilities under a focal hypothesis, unmatched by an adequate consideration of conditional probabilities under the alternatives. These behaviors might arguably originate from a “matching bias”, defined by Evans (1998) – in the domain of propositional reasoning – as a tendency to consider as relevant only the information whose lexical content matches that of a propositional rule to be tested. Villejoubert and Mandel (2002) conjectured that a similar matching tendency might affect some probabilistic judgments. Positive testing and the preference for testing features probable under the focal hypothesis, together, accounted for almost 60% of the variance of the responses by the participants in Experiment 4.

9. Conclusions

In tasks where previous knowledge and motivational factors are unlikely to play an important role, people’s hypothesis-testing strategies are far less variegated and sophisticated than those observed in more contextualized studies, such as those by Trope and his colleagues (1997; 2004), Skov, Sherman and their colleagues (1986, 1992), or Poletiek and Berndsen (2000). In the present tasks the context – being minimal – possibly gave participants no reason whatsoever for preferring to risk false positive errors instead of false negative errors, or vice-versa. As a consequence, participants’ individual preferences, in aggregated form, were indifferent to the symmetrical or asymmetrical properties of a question (whether or not they actually grasped them). Participants anchored their judgments to an easily accessible (Trope & Bassok, 1982) formally relevant feature – the diagnosticity of the question, as shown by the correlations in Experiments 1 to 3 – and to other easily accessible, but formally irrelevant, superficial features, such as positivity (all Experiments) or the probability of a feature under the focal hypothesis (Experiment 4). More complex, symmetry related features had little or no effect on responses, suggesting that hypothesis-testing behaviour in unfamiliar, abstract contexts is driven by simple, intuitive evaluations, more than complex, analytical ones.

In conclusion, in the light of our results, two main factors seem to drive people's testing preferences in abstract tasks: diagnosticity, and positivity. Contrary to some previous claims, extremity – either in its proper meaning of a tendency toward asymmetrically disconfirming testing, or in the more general meaning of asymmetric testing of both sorts – does not play a significant role. Asymmetric testing, of both types, certainly occurs sometimes; yet, the circumstances that allow to observe and replicate it are in need of being further qualified.

Acknowledgment

This study was partly funded by a PRIN 2006 grant to the first author by the Italian government, and by a 2007 FAR grant to the first author by the University of Milano-Bicocca. The authors thank Klaus Fiedler and an unknown reviewer for their helpful comments and suggestions, and Shanti Maria Utermark for proof reading the article.

References

- Baron, J., Beattie, J., & Hershey, J.C. (1988). Heuristics and biases in diagnostic reasoning. *Organizational Behavior and Human Decision Processes*, 42, 88-110.
- Beyth-Marom, R., & Fischhoff, B. (1983). Diagnosticity and pseudodiagnosticity. *Journal of Personality and Social Psychology*, 45, 6, 1185-1195.
- Cameron, J. A., & Trope, Y. (2004). Stereotype-biased Search and Processing of Information about Group Members. *Social Cognition*, 22, 6, 650-672.
- Cheng, P.W., & Holyoak, K.J. (1985), "Pragmatic reasoning schemas". In *Cognitive Psychology*, 17, pp. 391-416.
- Cherubini, P., Castelvechio, E., & Cherubini, A.M. (2004). Generation of hypotheses in Wason's 2-4-6 task: An information theory approach. *Quarterly Journal of Experimental Psychology*, 58, 309-332.
- Crupi, V., Tentori, K., & Gonzalez, M. (2007). On Bayesian theories of evidential support: Normative and descriptive considerations. *Philosophy of Science*, 74, 229-252..
- Evans, J. St. B. T. (1998). Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking and Reasoning*, 4, 45-82.
- Fiedler, K. (2000). Beware of samples! A Cognitive-ecological sampling theory of judgments biases. *Psychological Review*, 107, 659-676.
- Fiedler, K., Brinkmann, B., Betsch, T., & Wild, B. (2000). A sampling approach to biases in conditional probability judgments: Beyond base rate neglect and statistical format. *Journal of Experimental Psychology: General*, 129, 399-418.
- Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis Evaluation From a Bayesian Perspective. *Psychological Review*, 90, 3, 239-260.
- Gale, M., & Ball, L.J. (2006). Dual-Goal Facilitation in Wason's 2-4-6 Task: What Mediates Successful Rule Discovery? *Quarterly Journal of Experimental Psychology*, 59, 873-885.

- 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.
- Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive thought*. New York: Cambridge University Press.
- Good, I.J. (1950). *Probability and the weighing of evidence*. London: Charles Griffin & Company.
- Good, I.J. (1960). Weight of Evidence, corroboration, Explanatory power, Information and the Utility of Experiments. *Journal of the Royal Statistical Society. Series B (Methodological)*, 22, 319-331.
- Good, I.J. (1983). *Good thinking: the foundations of probability and its applications*. Minneapolis, MN: University of Minnesota Press.
- Hammond, K. R. (2007). *Beyond rationality: The search for wisdom in a troubled time*. Oxford: Oxford University Press.
- Henrion, M., & Fischhoff, B. (1986). Assessing uncertainty in physical constants. *American Journal of Physics*, 54, 791-798.
- Jones, M., & Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision*, 50, 59-99.
- Kareev, Y., Halberstadt, N., & Shafir, D. (1993). Improving performance and increasing the use of non-positive testing in a rule-discovery task. *Quarterly Journal of Experimental Psychology*, 46A, 729-742.
- Klayman, J. (1995). Varieties of confirmation bias. *The Psychology of Learning and Motivation*, 32, 385-418.
- Klayman, J., & Ha, Y.-W. (1987). Confirmation, Disconfirmation, and Information in Hypothesis Testing. *Psychological Review*, 94, 2, 211-228.
- Klayman, J., & Ha, Y.W. (1989). Hypothesis testing in rule discovery: Strategy, structure, and content. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 15, 596-604.

- Koehler, D. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes*, 56, 28-55.
- Lord, C.G., Ross, L., & Lepper, M.R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37, 11, 2098-2109.
- Mandel, D.R., & Lehman, D.R. (1998). Integration of contingency information in judgments of cause, covariation and probability. *Journal of Experimental Psychology: General*, 127, 269-285.
- McDonald, J. (1992). Is strong inference superior to simple inference? *Syntheses*, 92, 261-282.
- 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. London: Blackwell.
- Newman, J., Wolff, W.T., & Hearst E.J. (1980) The feature-positive effect in adult human subjects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 6, 630-50.
- Nickerson, R. S. (1996). Hempel's paradox and Wason's selection task: Logical and psychological puzzles of confirmation. *Thinking and Reasoning*, 2, 1-31.
- Nickerson, R.S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology*, 2, 175-220.
- Nisbett, R. E., & Ross, L. D. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.
- Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101, 608-631.
- Oaksford, M., & Chater, N. (2003). *Optimal data selection: Revision, review, and re-evaluation*. *Psychonomic Bulletin & Review*, 10, 289-318
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality: the probabilistic approach to human reasoning*. Oxford: Oxford University Press.

- Osteyee D.B., & Good I.J.(1974). *Information, weight of evidence, the singularity between probability measures and signal detection, lecture notes in mathematics*. Berlin/NewYork: Springer-Verlag.
- Poletiek, F.H., & Berndsen, M.T. (2000). Hypothesis Testing as Risk Behaviour with Regard to Beliefs. *Journal of Behavioural Decision Making*, 13, 107-123.
- Poses, R. M., Bekes, C., Copare, F. J., & Scott, W. E. (1990). What difference do two days make? The inertia of physicians' sequential prognostic judgments for critically ill patients. *Medical Decision Making*, 10, 6-14.
- Ross, L., & Anderson, C. (1982). Shortcomings in the attribution process: On the origins and maintenance of erroneous social assessments. In A. Tversky, D. Kahneman, & P. Slovic (Eds.), *Judgement Under Uncertainty: Heursitics and biases*. Cambridge, UK: Cambridge University Press.
- Rossi, S., Caverni, J.P., & Girotto, V. (2001). Hypothesis testing in a rule discovery problem: When a focused procedure is effective. *Quarterly Journal of Experimental Psychology*, 54, 263-267.
- 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.
- 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.
- Snyder, M., & Swann, W. B. (1978). Hypothesis-Testing Processes in Social Interaction. *Journal of Personality and Social Psychology*, 36, 11, 1202-1212.
- Spellman, B.A., Lopez, A., & Smith, E.E. (1999). Hypothesis testing: strategy selection for generalising versus limiting hypotheses. *Thinking and Reasoning*, 5, 67-91.
- Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition*, 57, 31-95.

- Trope, Y., & Bassok, M. (1982). Confirmatory and Diagnosing Strategies in Social Information Gathering. *Journal of Personality and Social Psychology*, 43, 1, 22-34.
- 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, 2, 229-241.
- Vallée-Tourangeau, F., Austin, N.G., & Ramkin, S. (1995). Inducing a rule in Wason's 2-4-6 task: A test of the information-quantity and goal-complementarity hypotheses. *Quarterly Journal of Experimental Psychology*, 48A, 895-914.
- 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.
- Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology I*. Harmondsworth, UK: Penguin.
- Wason, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12, 129-140.
- Wason, P.C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20, 273-281.
- Zuckerman, M., Knee, C.R., Hodgins, H.S., & Miyake, K.(1995). Hypothesis confirmation: the joint effect of positive test strategy and acquiescence response set. *Journal of Personality and Social Psychology*, 68, 52-60.

Appendices

Appendix A. A sample problem (problem 8, Experiments 1 and 2).

In front of you there are two decks of cards, each one composed of 100 cards. Printed on each card there are from zero to four geometric figures, chosen from among a triangle, a circle, a square, and a pentagon. The presence or absence of each figure on a card is fully independent from the presence or absence of any other figure. The following table shows the number of cards that display each figure:

	Triangle	Circle	Square	Pentagon
A deck	77	80	20	8
B deck	8	20	80	77

Now imagine that I draw a card at random from one of the two decks, but I don't tell you from which deck, and I don't show you the content of the card. Below there are four questions that you can ask, concerning the card. *Experiment 1 version:* Check the two questions that you deem most useful for determining whether the card was most likely drawn from the A deck (*B deck, for half participants*). *Experiment 2 version:* Rank in order the questions from most to least useful for determining whether the card was most likely drawn from the A deck (*B deck, for half participants*). Write "1" in the box beside the question or the questions that you surmise to be the most useful for determining whether the card was most likely drawn from the A deck; write "2" beside the question or the questions that you surmise to be second for usefulness, write "3" beside the question or questions third for usefulness, and so on. Remember to assign the same rank to the questions that you judge equally useful.

- ☐ Is there a triangle on the card?
- ☐ Is there a circle on the card?
- ☐ Is there a square on the card?
- ☐ Is there a pentagon on the card?

How confident are you in the correctness of your answer?

least confident 1 2 3 4 5 6 7 most confident

Note. In this problem the asymmetric questions are “Is there a triangle?” ($LR = 9.63$ for “yes”, $LR = 0.25$ for “no”) and “Is there a pentagon?” ($LR = 0.10$ for “yes”, and $LR = 4$ for “no”), whereas the other two questions are symmetric. The positive questions are “Is there a triangle?” and “Is there a circle?”, where the focal hypothesis is the A deck (otherwise the positive questions are the other two questions).

Appendix B. *The 8 problems in Experiments 1 and 2*

<i>Problem</i>	<i>Deck</i>	<i>p(triangle)</i>	<i>p(circle)</i>	<i>p(square)</i>	<i>p(pentagon)</i>
<i>1</i>	<i>A</i>	0.8	0.9	0.08	0.2
	<i>B</i>	0.2	0.08	0.9	0.8
<i>2</i>	<i>A</i>	0.8	0.46	0.05	0.2
	<i>B</i>	0.2	0.05	0.46	0.8
<i>3</i>	<i>A</i>	0.8	0.9	0.1	0.3
	<i>B</i>	0.3	0.1	0.9	0.8
<i>4</i>	<i>A</i>	0.95	0.8	0.2	0.54
	<i>B</i>	0.54	0.2	0.8	0.95
<i>5</i>	<i>A</i>	0.8	0.48	0.12	0.2
	<i>B</i>	0.2	0.12	0.48	0.8
<i>6</i>	<i>A</i>	0.8	0.88	0.52	0.2
	<i>B</i>	0.2	0.52	0.88	0.8
<i>7</i>	<i>A</i>	0.8	0.92	0.23	0.2
	<i>B</i>	0.2	0.23	0.92	0.8
<i>8</i>	<i>A</i>	0.77	0.8	0.2	0.08
	<i>B</i>	0.08	0.2	0.8	0.77

Appendix C. Some formal properties of the problems used in Experiments 1 and 2.

Problem	Question	Response	$p(\text{resp})$	I.G.	W.E.	Expected I.G.	Mean W.E.	Focus on Deck A	Focus on Deck B
1	Triangle?	Yes	.5	.28	6.02	.28	6.02	Symmetric, positive	Symmetric, negative
		No	.5	.28	6.02				
	Circle?	Yes	.49	.6	10.51	.56	10.07	Confirming, positive	Disconfirming, negative
		No	.51	.53	9.64				
	Square?	Yes	.49	.6	10.51	.56	10.07	Disconfirming, negative	Confirming, positive
		No	.51	.53	9.64				
2	Triangle?	Yes	.5	.28	6.02	.28	6.02	Symmetric, positive	Symmetric, negative
		No	.5	.28	6.02				
	Circle?	Yes	.26	.53	9.64	.18	4.29	Confirming, positive	Disconfirming, negative
		No	.75	.06	2.45				
	Square?	Yes	.26	.53	9.64	.18	4.29	Disconfirming, negative	Confirming, positive
		No	.75	.06	2.45				
3	Triangle?	Yes	.5	.28	6.02	.28	6.02	Symmetric, negative	Symmetric, positive
		No	.5	.28	6.02				
	Circle?	Yes	.55	.15	4.26	.19	4.79	Disconfirming, positive	Confirming, negative
		No	.45	.24	5.44				
	Square?	Yes	.5	.53	9.54	.53	9.54	Symmetric, positive	Symmetric, negative
		No	.5	.53	9.54				
4	Triangle?	Yes	.55	.16	4.26	.2	4.79	Confirming, negative	Disconfirming, positive
		No	.45	.24	5.44				
	Circle?	Yes	.75	.06	2.45	.18	4.29	Disconfirming, positive	Confirming, negative
		No	.26	.53	9.64				
	Square?	Yes	.5	.28	6.02	.28	6.02	Symmetric, positive	Symmetric, negative
		No	.5	.28	6.02				
5	Triangle?	Yes	.5	.28	6.02	.28	6.02	Symmetric, negative	Symmetric, positive
		No	.5	.28	6.02				
	Circle?	Yes	.3	.28	6.02	.12	3.41	Confirming, positive	Disconfirming, negative
		No	.7	.05	2.28				
	Square?	Yes	.3	.28	6.02	.12	3.41	Disconfirming, negative	Confirming, positive
		No	.7	.05	2.28				
6	Triangle?	Yes	.5	.28	6.02	.28	6.02	Symmetric, positive	Symmetric, negative
		No	.5	.28	6.02				
	Circle?	Yes	.3	.28	6.02	.12	3.41	Disconfirming, positive	Confirming, negative
		No	.7	.05	2.28				
	Square?	Yes	.3	.28	6.02	.12	3.41	Confirming, negative	Disconfirming, positive
		No	.7	.05	2.28				
7	Triangle?	Yes	.5	.28	6.02	.28	6.02	Symmetric, negative	Symmetric, positive
		No	.5	.28	6.02				
	Circle?	Yes	.58	.28	6.02	.4	7.64	Disconfirming, positive	Confirming, negative
		No	.43	.56	9.83				
	Square?	Yes	.58	.28	6.02	.4	7.64	Confirming, negative	Disconfirming, positive
		No	.43	.56	9.83				
8	Triangle?	Yes	.5	.28	6.02	.28	6.02	Symmetric, positive	Symmetric, negative
		No	.5	.28	6.02				
	Circle?	Yes	.43	.55	9.83	.39	7.64	Confirming, positive	Disconfirming, negative
		No	.58	.28	6.02				
	Square?	Yes	.5	.28	6.02	.28	6.02	Symmetric, negative	Symmetric, positive
		No	.5	.28	6.02				
	Pentagon?	Yes	.43	.56	9.83	.4	7.64	Disconfirming, negative	Confirming, positive
		No	.58	.28	6.02				

$p(\text{resp})$ = probability of receiving that response

I.G.(information gain) = $p(H)\log_2[1/p(H)] + p(-H)\log_2[1/p(-H)] - p(H|E)\log_2[1/p(H|E)] + p(-H|E)\log_2[1/p(-H|E)]$; this is the difference in Shannon's entropy after a response (E) is received, in bits.

W.E. (weight of evidence) = $10 \cdot \log_{10} LR$, in decibans (Good, 1983). This is often used instead of the raw LR.

Expected I.G. = $p(\text{yes}) \cdot I.G.(\text{yes}) + p(\text{no}) \cdot I.G.(\text{no})$. A measure of the diagnosticity of a question

Mean W.E. = $p(\text{yes}) \cdot W.E.(\text{yes}) + p(\text{no}) \cdot W.E.(\text{no})$. Another measure of the diagnosticity of a question.

The last two columns report the classification of each question, depending on the focal hypothesis (either Deck A or B)

Appendix D. The 8 problems used in Experiment 3.

<i>Problem</i>	<i>Deck</i>	<i>p(triangle)</i>	<i>p(circle)</i>	<i>p(square)</i>	<i>p(pentagon)</i>
<i>1</i>	<i>A</i>	0.82	0.92	0.06	0.18
	<i>B</i>	0.18	0.06	0.92	0.82
<i>2</i>	<i>A</i>	0.81	0.47	0.04	0.19
	<i>B</i>	0.19	0.04	0.47	0.81
<i>3</i>	<i>A</i>	0.81	0.88	0.12	0.39
	<i>B</i>	0.39	0.12	0.88	0.81
<i>4</i>	<i>A</i>	0.96	0.86	0.14	0.37
	<i>B</i>	0.37	0.14	0.86	0.96
<i>5</i>	<i>A</i>	0.82	0.41	0.09	0.18
	<i>B</i>	0.18	0.09	0.41	0.82
<i>6</i>	<i>A</i>	0.84	0.92	0.58	0.16
	<i>B</i>	0.16	0.58	0.92	0.84
<i>7</i>	<i>A</i>	0.79	0.94	0.25	0.21
	<i>B</i>	0.21	0.25	0.94	0.79
<i>8</i>	<i>A</i>	0.75	0.79	0.21	0.06
	<i>B</i>	0.06	0.21	0.79	0.75

Appendix E. *Some formal properties of the 8 problems in Experiment 3 (legend in appendix C).*

Problem	Question	Response	$p(\text{resp})$	I.G.	W.E.	Expected I.G.	Mean W.E.	Focus on Deck A	Focus on Deck B
1	Triangle?	Yes	.5	.32	6.59	.32	6.59	Symmetric, positive	Symmetric, negative
		No	.5	.32	6.59				
	Circle?	Yes	.49	.67	11.86	.63	11.27	Confirming, positive	Disconfirming, negative
		No	.51	.6	10.7				
	Square?	Yes	.49	.67	11.86	.63	11.27	Disconfirming, negative	Confirming, positive
		No	.51	.6	10.7				
2	Pentagon?	Yes	.5	.32	6.59	.32	6.59	Symmetric, negative	Symmetric, positive
		No	.5	.32	6.59				
	Triangle?	Yes	.5	.3	6.3	.3	6.3	Symmetric, positive	Symmetric, negative
		No	.5	.3	6.3				
	Circle?	Yes	.255	.6	10.7	.2	4.65	Confirming, positive	Disconfirming, negative
		No	.745	.06	2.58				
3	Square?	Yes	.255	.6	10.7	.2	4.65	Disconfirming, negative	Confirming, positive
		No	.745	.06	2.58				
	Pentagon?	Yes	.5	.3	6.3	.3	6.3	Symmetric, negative	Symmetric, positive
		No	.5	.3	6.3				
	Triangle?	Yes	.6	.09	3.17	.14	3.93	Disconfirming, positive	Confirming, negative
		No	.4	.20	5.07				
4	Circle?	Yes	.5	.47	8.65	.47	8.65	Symmetric, positive	Symmetric, negative
		No	.5	.47	8.65				
	Square?	Yes	.5	.47	8.65	.47	8.65	Symmetric, negative	Symmetric, positive
		No	.5	.47	8.65				
	Pentagon?	Yes	.6	.09	3.17	.14	3.93	Confirming, negative	Disconfirming, positive
		No	.4	.20	5.07				
5	Triangle?	Yes	.665	.15	4.14	.32	6.76	Disconfirming, positive	Confirming, negative
		No	.335	.67	11.97				
	Circle?	Yes	.5	.42	7.88	.42	7.88	Symmetric, positive	Symmetric, negative
		No	.5	.42	7.88				
	Square?	Yes	.5	.42	7.88	.42	7.88	Symmetric, negative	Symmetric, positive
		No	.5	.42	7.88				
6	Pentagon?	Yes	.665	.14	4.14	.32	6.76	Confirming, negative	Disconfirming, positive
		No	.335	.67	11.97				
	Triangle?	Yes	.5	.32	6.59	.32	6.59	Symmetric, positive	Symmetric, negative
		No	.5	.32	6.59				
	Circle?	Yes	.25	.32	6.59	.11	3.06	Confirming, positive	Disconfirming, negative
		No	.75	.04	1.88				
7	Square?	Yes	.25	.32	6.59	.11	3.06	Disconfirming, negative	Confirming, positive
		No	.75	.04	1.88				
	Pentagon?	Yes	.5	.32	6.59	.32	6.59	Symmetric, negative	Symmetric, positive
		No	.5	.32	6.59				
	Triangle?	Yes	.5	.37	7.2	.37	7.2	Symmetric, positive	Symmetric, negative
		No	.5	.37	7.2				
8	Circle?	Yes	.75	.04	2	.12	3.3	Disconfirming, positive	Confirming, negative
		No	.25	.37	7.2				
	Square?	Yes	.75	.04	2	.12	3.3	Confirming, negative	Disconfirming, positive
		No	.25	.37	7.2				
	Pentagon?	Yes	.5	.37	7.2	.37	7.2	Symmetric, negative	Symmetric, positive
		No	.5	.37	7.2				
9	Triangle?	Yes	.5	.26	5.75	.26	5.75	Symmetric, positive	Symmetric, negative
		No	.5	.26	5.75				
	Circle?	Yes	.595	.26	5.75	.41	7.86	Disconfirming, positive	Confirming, negative
		No	.405	.63	10.97				
	Square?	Yes	.595	.26	5.75	.41	7.86	Confirming, negative	Disconfirming, positive
		No	.405	.63	10.97				
10	Pentagon?	Yes	.5	.26	5.75	.26	5.75	Symmetric, negative	Symmetric, positive
		No	.5	.26	5.75				
	Triangle?	Yes	.405	.62	10.97	.40	7.86	Confirming, positive	Disconfirming, negative
		No	.595	.26	5.75				
	Circle?	Yes	.5	.26	5.75	.26	5.75	Symmetric, positive	Symmetric, negative
		No	.5	.26	5.75				
11	Square?	Yes	.5	.26	5.75	.26	5.75	Symmetric, negative	Symmetric, positive
		No	.5	.26	5.75				
	Pentagon?	Yes	.405	.63	10.97	.41	7.86	Disconfirming, negative	Confirming, positive
		No	.595	.26	5.75				
	Triangle?	Yes	.405	.62	10.97	.40	7.86	Confirming, positive	Disconfirming, negative
		No	.595	.26	5.75				

Appendix F. *The 8 problems used in Experiment 4.*

Problem	Deck	P(triangle)	P(circle)	P(square)	P(pentagon)	Diagnosticity of questions
1	A	0.11	0.98	0.89	0.02	Low (mean WE = .87, I.G. between .026 and .008)
	B	0.02	0.89	0.98	0.11	
2	A	0.68	0.11	0.58	0.02	
	B	0.58	0.02	0.68	0.11	
3	A	0.98	0.68	0.89	0.58	
	B	0.89	0.58	0.98	0.68	
4	A	0.68	0.49	0.58	0.39	
	B	0.58	0.39	0.68	0.49	
5	A	0.98	0.38	0.02	0.62	High (mean W.E. = 4.15, I.G. between .15 and .17)
	B	0.62	0.02	0.38	0.98	
6	A	0.38	0.62	0.02	0.18	
	B	0.02	0.18	0.38	0.62	
7	A	0.62	0.98	0.18	0.62	
	B	0.18	0.62	0.62	0.98	
8	A	0.62	0.88	0.45	0.18	
	B	0.18	0.45	0.88	0.62	

Notes: Bold numerals identify which probability is extreme in each question (i.e., further removed from .5 with respect to the other hypothesis). In the design, extreme probabilities were classified into:

Very high: $p > 0.9$

High: $.5 < p < 0.9$

Low: $.5 > p > 0.1$

Very low: $p < 0.1$

Their association either to the focal or non focal hypothesis generated negative disconfirming, negative confirming, positive disconfirming, or positive confirming questions (see Appendix G).

Appendix G. Some formal properties of the problems in Experiment 4.

Problem	Question	Response	$p(\text{resp})$	I.G.	W.E.	Focus on Deck A	Focus on Deck B
1	Triangle?	Yes	.065	.39	7.40	Confirming, positive	Disconfirming, negative
		No	.935	.001	.42		
	Circle?	Yes	.935	.001	.42	Disconfirming, positive	Confirming, negative
		No	.065	.39	7.40		
	Square?	Yes	.935	.001	.42	Confirming, negative	Disconfirming, positive
		No	.065	.39	7.40		
	Pentagon?	Yes	.065	.39	7.40	Disconfirming, negative	Confirming, positive
		No	.935	.001	.42		
2	Triangle?	Yes	.63	.005	.69	Disconfirming, positive	Confirming, negative
		No	.37	.015	1.18		
	Circle?	Yes	.065	.39	7.40	Confirming, positive	Disconfirming, negative
		No	.935	.001	.42		
	Square?	Yes	.63	.005	.69	Confirming, negative	Disconfirming, positive
		No	.37	.015	1.18		
	Pentagon?	Yes	.065	.39	7.40	Disconfirming, negative	Confirming, positive
		No	.935	.001	.42		
3	Triangle?	Yes	.935	.001	.42	Disconfirming, positive	Confirming, negative
		No	.065	.39	7.4		
	Circle?	Yes	.63	.005	.69	Disconfirming, positive	Confirming, negative
		No	.37	.014	1.18		
	Square?	Yes	.935	.001	.42	Confirming, negative	Disconfirming, positive
		No	.065	.39	7.4		
	Pentagon?	Yes	.63	.005	.69	Confirming, negative	Disconfirming, positive
		No	.37	.014	1.18		
4	Triangle?	Yes	.63	.005	.69	Disconfirming, positive	Confirming, negative
		No	.37	.014	1.18		
	Circle?	Yes	.44	.01	.99	Confirming, positive	Disconfirming, negative
		No	.56	.005	.78		
	Square?	Yes	.63	.005	.69	Confirming, negative	Disconfirming, positive
		No	.37	.014	1.18		
	Pentagon?	Yes	.44	.01	.99	Disconfirming, negative	Confirming, positive
		No	.56	.005	.78		
5	Triangle?	Yes	.8	.035	1.99	Disconfirming, positive	Confirming, negative
		No	.2	.714	12.79		
	Circle?	Yes	.2	.715	12.79	Confirming, positive	Disconfirming, negative
		No	.8	.035	1.99		
	Square?	Yes	.2	.714	12.79	Disconfirming, negative	Confirming, positive
		No	.8	.035	1.99		
	Pentagon?	Yes	.8	.035	1.99	Confirming, negative	Disconfirming, Positive
		No	.2	.714	12.79		
6	Triangle?	Yes	.2	.714	12.79	Confirming, positive	Disconfirming, negative
		No	.8	.035	1.99		
	Circle?	Yes	.4	.233	5.37	Confirming, positive	Disconfirming, negative
		No	.6	.096	3.34		
	Square?	Yes	.2	.714	12.79	Disconfirming, negative	Confirming, positive
		No	.8	.035	1.99		
	Pentagon?	Yes	.4	.233	5.37	Disconfirming, negative	Confirming, positive
		No	.6	.096	3.34		
7	Triangle?	Yes	.4	.233	5.37	Confirming, positive	Disconfirming, negative
		No	.6	.096	3.34		
	Circle?	Yes	.8	.035	1.99	Disconfirming, positive	Confirming, negative
		No	.2	.714	12.79		
	Square?	Yes	.4	.233	5.37	Disconfirming, negative	Confirming, positive
		No	.6	.096	3.34		
	Pentagon?	Yes	.8	.035	1.99	Confirming, negative	Disconfirming, positive
		No	.2	.714	12.79		
8	Triangle?	Yes	.4	.233	5.37	Confirming, positive	Disconfirming, negative
		No	.6	.096	3.34		
	Circle?	Yes	.665	.075	2.91	Disconfirming, positive	Confirming, negative
		No	.335	.32	6.61		
	Square?	Yes	.665	.075	2.91	Confirming, negative	Disconfirming, positive
		No	.335	.32	6.61		
	Pentagon?	Yes	.4	.233	5.37	Disconfirming, negative	Confirming, positive
		No	.6	.096	3.34		

For the expected I.G. and mean W.E. of questions, see appendix F.

Negative disconfirming questions: extreme in the “low” or “very low” ranges associated to the focal hypothesis.

Positive confirming questions: extreme in the “low” or “very low” ranges associated to the non-focal hypothesis.

Negative confirming questions: extreme in the “high” or “very high” ranges associated to the non-focal hypothesis.

Positive disconfirming questions: extreme in the “high” or “very high” ranges associated to the focal hypothesis.