UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Search and Evaluation Strategies in Belief Revision: Psychological Mechanisms and Normative Deviations

by

Patrice Rusconi

Advisor: Prof. Paolo Cherubini

Committee in charge:

Prof. Norma De Piccoli, Università degli Studi di Torino

Prof. Renata Tambelli, Università degli Studi di Roma "La Sapienza"

Prof. Katya Tentori, Università degli Studi di Trento - polo di Rovereto

2010

Chapter 2 consists of a reprint of the following published article, in which the dissertation author was second author and co-investigator:

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.

To my parents and my sister

Table of contents

Acknowledgements

I am grateful to many people:
to my parents and my sister, for support, encouragements, suggestions, patience and love;
to Tiziana Mazzotta and Marco Mion, my best friends who have never ceased to encourage, support and give me suggestions;
to Craig McKenzie and David Huber who hosted me in their labs at the University of California, San Diego;
to Shlomi Sher, whom I thank in particular for the many passionate and smart discussions;
to Luciano Giromini, for invaluable help;
to Marco D'Addario, Paolo Riva, Simona Sacchi, Selena Russo, Marco Brambilla, Carlo Geraci, Davide Crepaldi, Marco Marelli, Luca Andrighetto, Chiara Colombo, Stefano Ravasio, Tomas Borgogna, Emanuela Brusadelli, Laura Bonalume, Paola Riva and all my colleagues in the PhD course;
to Samantha Callahan and Alvaro Brander and all the guys I met in San Diego;
to all the guys with whom I have worked and I'm still working at the "Giornale di Lecco".
Finally, special thanks to all of the participants to our experiments, many of whom volunteered to participate in our studies. I appreciate the time they have spent for us as well as their comments on our studies.

Abstract

The procedures people adopt in order to seek out and use information have been the focus of empirical research since long in psychology, especially so from the late 50s. This dissertation addresses some key questions left unanswered by a series of seminal studies which date back to the 80s and the 90s on information-gathering and information-use strategies.

We first dealt with the question-asking preferences that people exhibit in an abstract task of hypothesis testing. Specifically, we pitted against one another the tendencies to ask positive questions, wherein the confirming answer is expected given the truth of the working hypothesis, and to pose asymmetric queries, wherein the anticipated outcomes of a dichotomous question (i.e., “yes” and “no” answers) convey different amounts of information. Finally, we investigated whether or not people prefer either asymmetrically confirming queries (i.e., questions for which the confirming answer weights more than the disconfirming answer) or asymmetrically disconfirming queries (i.e., questions for which the disconfirming answer conveys more information than the confirming answer). We found a robust tendency to ask positive testing, in keeping with the literature, but neither a preference for asymmetric questions, nor a predominant use of symmetric testing. Furthermore, we showed, correlationally, that people are sensitive to the diagnosticity of questions, as some previous studies in the literature pointed out. Finally, it emerged an interaction between the positivity of questions and the confirming valence of asymmetric queries. A close analysis of the latter finding allowed us to undermine the possibility that people would try to maximize the probability of occurrence of the tested feature, while suggesting a less sophisticated strategy based on the consideration of an easily accessible feature, that is, the probability of a feature under the working hypothesis.

After further deepening the study of strategies adopted in the testing phase of hypothesis development, we turned to the evaluation stage. Specifically, we addressed the finding emerged in previous studies of the relative insensitivity of people to the different diagnosticity conveyed by different answers (i.e., “yes” and “no”) to the same question in an abstract task. We showed that not only people might exhibit insensitivity but also

oversensitivity to differentially informative answers, indicating a more general failure in information use than previously thought. We also addressed the issue of why people are either insensitive or oversensitive to answer diagnosticity. We provided evidence that an explanation based on the use of the feature-difference heuristic, which has been proposed previously in the literature and wherein people's estimates are influenced by the difference between the likelihoods, seems unable to explain people's behavior. By contrast, we found that people prefer to rely on an averaging strategies, in particular on the average between the prior probability and the likelihood.

Finally, we investigated an aspect emerged but not directly investigated by previous studies on hypothesis evaluation, that is the feature-positive effect, wherein people tend to overestimate the presence of a feature as opposed to its absence. The results of three experiments with abstract tasks strongly confirmed that the hypothesized effect influences both frequency and accuracy of participants' responses. We also found that participants exhibited some sensitivity to the formal amount of information, although only with respect to the present clues.

Overall, the series of experiments presented in this dissertation contributes to better clarify how people search for information, by showing that they might rely both on formally relevant and formally irrelevant properties of the information they have at hand and by putting into question the alleged tendency to hypothesis confirmation, defined as a maximization of the probability of a confirming datum. Furthermore, these experiments help understand how people treat information, by specifying how people misweigh differentially diagnostic answers and showing that a psychologically compelling tendency, namely the feature-positive effect, might, at least in part, account for people's information use.

Introduction

In daily life we are called upon to make conjectures about the state of the world, test them by looking for new pieces of information and either confirm or revise our initial opinions in light of the new collected evidence in order to undertake an action. Often these cognitive processes are not achieved exactly in this order, instead people might loop back at various stages of this inferential cycle or skip some of them (e.g., Trope & Liberman, 1996). Furthermore, all of these activities might be explicit to different extents, inasmuch some hypotheses might be more accessible than others (e.g., social stereotypes), some data might be more psychologically compelling than others (e.g., certain pattern of symptoms might shift physicians' confidence toward a hypothesized disease more readily than others, Cherubini, Russo, Rusconi, D'Addario, & Boccuti, 2009), some processes might be achieved automatically and some others require, instead, a more analytical elaboration (e.g., handling negative, as opposed to positive, information, e.g., Hearst, 1991; Van Wallendael, 1995; Wason, 1959, 1961). Regardless, generating, testing and evaluating hypotheses are activities which people engage in to cope with their physical and social environment (e.g., Poletiek, 2001), often dealing with fallible information and limited resources. The main aim of these activities is reducing uncertainty ("removing the doubt" which is at the base of thinking, Baron, 2008, p. 7). The ultimate consequence of developing hypotheses bears on undertaking the most appropriate course of action given a certain context and the constraints of the environment.

The present set of studies aimed to scratch the surface of the psychological mechanisms which influence hypothesis testing and hypothesis evaluation. We will go through a series of experiments using probabilistic tasks with abstract material whose results might help better understand the strategies recruited by participants when searching for and interpreting information.

Chapter 1 provides, in a succinct fashion, some basic principles and background literature which are referred to in the experiments we shall describe in the subsequent chapters, as well as the framework in which they are embodied. In Chapter 2, question-asking preferences of participants in abstract probabilistic tasks are concerned.

Specifically, in four experiments we compared the relative influence of positive and a/symmetric testing (as well as the confirming vs. disconfirming value of asymmetric tests) when choosing which of four questions is the most useful to test the validity of one of two mutually exclusive and jointly exhaustive hypotheses. Once addressed the issue of hypothesis *testing*, we shall turn to the *evaluation* strategies (Chapter 3), showing, through two experiments, that the relative insensitivity to the differential diagnosticity of different answers to the same question found in previous studies (McKenzie, 2006; Skov & Sherman, 1986; Slowiaczek, Klayman, Sherman, & Skov, 1992) might turn to *oversensitivity* under some circumstances, thus indicating a more general failure in information use than previously claimed. We shall also provide tentative evidence about which model can best account for the revision of initial opinions in light of different answers to a dichotomous question. In Chapter 4, we shall present three studies in which we investigated whether and to what extent a well-known psychological mechanism, the focusing on presence as opposed to absence (i.e., feature-positive effect), affects the evaluation of two mutually exclusive and jointly exhaustive hypotheses. Finally, we shall draw the conclusions.

All of the experiments we are going to present are with abstract materials. This is meant to show how the cognitive mechanisms involved in hypothesis testing and evaluation work at the net of motivational factors and previous, strongly-held beliefs. In other words, we want to describe what is at the *core* of a variety of contextualized inferential activities, with which, either explicitly or implicitly, we are familiar because they fill up our lives. However, the reader who is interested in the more applicative side of the matter is referred to other studies which we conducted during the past few years in the fields of impression formation (Brambilla, Rusconi, Sacchi, & Cherubini, in press, Study 2), social hypothesis testing (Sacchi, Rusconi, Russo, Bettiga, & Cherubini, accepted with minor revision) and health psychology (Riva, Rusconi, Montali, & Cherubini, in press; Rusconi, Riva, Cherubini, & Montali, 2010).

Chapter 1

Testing and evaluating hypotheses

Imagine to be a personal investigator who is asked to find a missing woman by her husband. At the phone, he tells you that she has just taken off from a private residential detoxification center where she was sent to get off drugs. He tells you that he does not have any idea of where she could be at the moment and he is worried because she is suicidable. You check the last movements of her credit card, even though her husband told you that her card has been blocked due to her drug addiction. You find out that someone bought a bus ticket with that card the day before. You thus decide to call some friends of the woman to check her husband's statements and to acquire more information about her emotional state, her habits and any possible event in her life in the last few weeks. You get to know that the couple was in crisis in the last period and she was thinking of breaking up and divorcing. Her best friend tells you that she is very rich and she has a house in the countryside, where she used to go on vacation, so you hypothesize that she might be there and you head for her house. During your trip you start thinking as to why her husband was so worried about her suicide tendencies, while none of her closest friends you heard mentioned this point. Once arrived, you enter her house and you start looking for her. It seems like there is anybody in. However, you keep looking for any clues, because the gate was unlocked, hence some intruders might have been or still be in. Eventually, you find out her body lying on the floor of her bedroom, with a gun and a note just nearby. You call the police. Apparently, there is no doubt: to be sure, it is suicide according to the police. However, you do not want to rush to judgment and you decide to wait for the lab results about all of the evidence that has been collected at the crime scene (e.g., gunshot residue, drugs, fingerprints, direction of blood flow, bullet holes, suicide note, time of death, etc.) before sharing the police's conclusion about this death. The husband's claims about her credit card (did he really believe it was blocked or was he lying?) and her emotional state (why her closest friends did not mention that she was suicidable?) point to some inconsistencies to you. Furthermore, it seems to you odd that the door was open. You definitely believe that the lab results are in need before drawing any strong conclusions. Once you receive the test results you are even more doubtful than before about the suicide hypothesis. The amount of heroin found in the body (1500 mg,

that is, more than three times higher than the lethal dose of heroin, which ranges from 200 to 500 mg) seem to be inconsistent with the possibility of pulling the trigger. Furthermore, the fingerprints on the shells are not legible. Finally, the handwriting of the suicide note might not belong to the victim. Is it possible that a person who has developed tolerance to heroin can survive that dose and be able to handle a shotgun and pull the trigger? Is the awkward handwriting a direct consequence of the heroin injection, again assuming that the addict could still be able to jot down a few lines after injecting such a large dose? Taken together, the available pieces of evidence shift your belief from a plausible suicide case to an undetermined death. You decide that further investigations are in need to get a clearer picture of this case.

Crime investigations, as the one sketched above, are but one of the many situations in which people raise questions about others (in the example we gave, "why her closest friends did not mention that she was suicidable?") or ourselves, make conjectures ("some intruders might have been or still be in"), generate hypotheses (the missing woman could be at her countryside house), look for information corroborating or disconfirming their initial beliefs (lab tests), interpret the evidence gathered from the environment or retrieved from their memory (the huge dose of heroin in her body, the illegible fingerprints, the handwriting are clues which put into question the hypothesis of a suicide), and they draw conclusions which then will likely influence or determine their choices and behaviors (it might be a murder despite the crime scene leads to think of a suicide, thus further investigations are in need). All these activities, in various guises, are common not only to professions, such as that of private investigator, but also to mundane situations, spanning from impression formation to scientific testing of theories, from visual perception to medical diagnosis, from getting acquainted to the formation and maintenance of stereotypes, from language acquisition to problem solving (e.g., Baron, Beattie, & Hershey, 1988; Cherubini, 2007; Evett, Devine, Hirt, & Price, 1994; Fiedler & Walther, 2004; McKenzie, 2004; Poletiek, 2001). Indeed, in many aspects of our lives we rely on these different processes which constitute hypothesis development (Klayman, 1995), wherein people look for a correspondence between the beliefs they hold and the actual state of affairs (McKenzie, 2004; a.k.a., "correspondence competence", Hammond, 2007). As stated by Craig McKenzie (2004): "We all engage in hypothesis development on a regular basis as we try to organize and impose structure on our complex world".

The multiple steps of this process can be reduced to three main stages (McKenzie, 2004): hypothesis generation, testing and evaluation (for a more fine-grained breakdown see Trope & Liberman, 1986). Hypothesis *generation* consists of figuring out a possible explanation or anticipating the outcome of a state of the world. The *testing* phase entails the assessment, through the collection of new information, of the working hypothesis in order to prove whether it is appropriate or not. The gathered evidence is then weighed and interpreted in the *evaluation* stage in order to refute, validate or modify the initial opinions (McKenzie, 2004).

The present work focuses on the testing and evaluation phases of hypothesis development, shedding light on how people seek out for evidence and how they interpret incoming data. Specifically, we will concentrate on the strategies used by people engaged in abstract probabilistic tasks and on the psychological mechanisms which might guide information gathering and information processing, showing how and to what extent people are driven by properties which are either formally relevant (e.g., the diagnostic value of a question) or formally irrelevant (e.g., the magnitude of the probability, under the working hypothesis, of a feature to inquire about or the presence, as opposed to the absence, of a clue to evaluate in order to select which of two jointly exhaustive and mutually exclusive hypotheses is the most plausible). This is meant to provide further explanations of people's behaviors in such tasks, giving tentative answers to questions such as: Do people's question-asking preferences lead to confirmation bias? Why are people relative insensitive to the differential diagnosticity of different answers to the same question? Both when testing and evaluating evidence, are people sensitive to the formal amount of information conveyed by the sought or given data?

The experiments we are going to present here are embodied in a probabilistic framework. Although people's judgments do not necessarily rely on quantifications, using probabilities is not arbitrary, because the world in which people make their judgments is characterized by uncertainty, to the point that "there are very few circumstances where such uncertainty is absent" (Hammond, 2007). And since "probabilities quantify uncertainty" (Edwards, 1968), probability theory seems to be well-suited to model human behavior (e.g., Oaksford & Chater, 2007). Specifically, we will refer to Bayes' theorem, a widely used formal criterion in inductive-reasoning studies since in 1963 Ward Edwards "introduced psychologists to Bayesian thinking with his paper *Bayesian Statistical Inference for Psychological Research* published in *Psychological Review*" (Newell,

2009). The Bayes' theorem derives from the axioms of classical probability theory, and thus its validity is uncontroversial (e.g., Cherubini, 2007; Edwards, 1968). However, the definition of probability entailed by the Bayes' theorem is more debated, because it differs from classic *aleatory* probability, which applies to games of chance and wherein the probability is conceived as the ratio of the number of favourable outcomes to the number of possible outcomes[1]. By contrast, according to the Bayesian approach, probability is still a number between 0 and 1, but it expresses the subjective *degrees of belief* with respect not only to a series of repeated events, but also to a single, unrepeated event. In other words, Bayesian probability is *epistemic*. Bayes' rule prescribes how people should revise their initial opinions in light of new information and can be expressed in terms of odds and likelihood ratios[2] (Beyth-Marom & Fischhoff, 1983; Fischhoff & Beyth-Marom, 1983), as follows:

$$[p\,(H \mid E) / p\,(\neg H \mid E)] = [p\,(H) / p\,(\neg H)] \times [p\,(E \mid H) / p\,(E \mid \neg H)]$$

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" (hereafter, LR): The ratio of the probability of observing "E", assuming the truth of "H", to the probability of nevertheless observing "E" if "H" were false.

The LR is a measure of the strength of confirmation (or falsification) conveyed by "E". It conveys an immediate and direct description of the impact of evidence on the revision of the initial belief. If it is 1, "E" does not change the probability of "H", and thus it is uninformative. If LR > 1, "E" confirms "H", by increasing its probability. If LR < 1,

[1] This principle has been first expressed in *Liber de Ludo Aleae* by the Italian physician and mathematician Girolamo Cardano (1501-1576), the first to attempt to provide a mathematical account of random phenomena (Tabak, 2004).

[2] This description of Bayes' rule is adapted from Cherubini, Rusconi, Russo, & Crippa (submitted). Missing the dog that failed to bark in the nighttime: On the overestimation of occurrences over non-occurrences in hypothesis testing.

"E" falsifies "H", by decreasing its probability (or, correspondingly, it confirms "¬H" by the magnitude of 1/LR).

First studies on hypothesis development

A Study of Thinking by Jerome S. Bruner, Jacqueline J. Goodnow and the late George A. Austin (1956) is usually referred to as a milestone in the empirical research on hypothesis development (e.g., McKenzie, 2004; Poletiek, 2001). It was the product of a several years program, started in 1951 at the Institute for Advanced Study in Princeton, which took into account the "mediation" of cognitive processes between stimuli and responses, thus going beyond the behaviorist approach and taking part to the "revival", as the authors stated in the preface of the book, of studies on higher mental functions which were of prominent interest in the years before the first World War. Specifically, the authors addressed the issues of "concept attainment" and "concept utilization" (Bruner et al., 1956), that is, processes which are at the basis of all inferential activities. For this reason, this work provided insights to a wealth of studies on hypothesis development which have flourished later on. The typical procedure used by the authors to investigate concept attainment (i.e., figuring out what a concept might be) made use of an array of instances (81 cards) which comprised four attributes and each could assume three values: shape (circle, square, cross), color (red, green, black), number (one figure, two figures, three figures), borders (one, two, three). Participants were told what was meant for "concept" and that the experimenter had a concept in mind which was illustrated by some of the cards in the array, but not by some others. They were required to determine what the concept was. At the beginning, the experimenter presented to participants a card which was a positive example of the concept. The participants had to choose, one at a time, a card which she/he deemed an appropriate instance of the concept. After any choice the participants were told whether or not the card was a positive example of the concept and they could hypothesize what the concept was. They were asked to attain the concept in the most efficient fashion.

Bruner et al. (1956) described four different selection strategies to which people can resort in this task: the simultaneous-scanning strategy, the successive-scanning strategy, the conservative-focussing strategy, and the focus-gambling strategy. The simultaneous scanning allows to get the maximum informativeness because a tester, after each choice of an instance, rules out some hypotheses while keeping others. However, this is a quite effortful procedure, because the seeker should keep in memory several independent

hypotheses and choose the next instances based on the elimination of as many as possible hypotheses for each chosen instance. By contrast, successive scanning reduces cognitive effort, in that it requires to take into account only one hypothesis at a time, though not guaranteeing to obtain the maximum information possible. Once a person has chosen a hypothesis to test, she/he looks for positive instances of the hypothesized concept, "the typical successive scanner then *limits his choices to those instances that provide a direct test of his hypothesis*" (Bruner et al., 1956, p. 85). Conservative focussing entails a focus of the tester on a positive instance and subsequent choices of instances in which there is a variation of but one attribute value of the focus card. If the change yields a positive instance then the modified attribute value will not be considered as part of the concept. Vice versa, if the change yields a negative instance, then the original attribute value will be retain as an attribute value of the concept. This strategy is conservative in that it guarantees that some information, though not the maximum possible, will be found. Since it allows to "*test the relevance of attributes*" (Bruner et al., 1956, p. 88), conservative focussing provides a cognitive economical way of eliminating hypotheses. Focus gambling differs from conservative focussing because the positive instance taken as a focus is changed in "*more than one* attribute value at a time" (Bruner et al., 1956, p. 89). This strategy entails a more risky approach than that of conservative focussing, because one may either reduce the steps leading to the correct concept, increase them, or in between. Specifically, if, for instance, one takes as a focus a positive card and then changes three attribute values of the focus card and still obtain a positive instance of the concept, then the attribute value shared by both the focus card and the changed card is the correct concept. However, it might be that the change in more than attribute value of the positive focus card yields a negative instance. In this case focus gambling is no more faster than conservative focussing. Indeed, one has to resort to simultaneous scanning to attain the concept, thus shifting from a search for relevant attributes to a test of appropriate hypotheses. Alternatively, the seeker has to decide not to use the information conveyed by the negative instance, in which case she/he might then choose instances which contain some of the information conveyed by that negative instance. In this sense, focus gambling, as opposed to conservative focussing, does not guarantee that each choice carries new evidence.

In their description Bruner et al. (1956) referred to the "potential informational value" and "*maximum* informativeness" of sought instances, the "positive" and "negative"

status of instances, the use of "direct" and "indirect" tests, and the "risk-regulating nature of a strategy" when choosing the instances to determine what the concept is. Diagnosticity (e.g., Trope & Bassok, 1982), positivity (e.g., Klayman & Ha, 1987), and asymmetry (e.g., Trope & Thompson, 1997, more specifically, see the risk-taking approach to hypothesis testing, Poletiek & Berndsen, 2000, and the error-avoidance interpretation of hypothesis testing given by Trope & Liberman, 1996), which subsequent studies suggested to play key roles in driving people's preferences in hypothesis testing (e.g., McKenzie, 2004), resemble the concepts that Bruner and co-authors described in the 50s.

Other seminal studies on how people make conjectures and test them has been conducted by Peter Wason (1924-2003) in the 1960s. The influential British psychologist devised two experimental paradigms which were subsequently reproduced in a wealth of studies in the last four decades, namely the "2−4−6" or "rule discovery" task (1960) and the "selection task" (1966).

The 2−4−6 task was devised as a further development of the Bruner et al.'s (1956) work on concept attainment strategies, showing that people might interpret instances which can *sufficiently* confirm a hypothesis as *necessary-and-sufficient* evidence of the hypothesized concept. Participants were told that the triplet 2−4−6 conformed to a rule which they were to discover (i.e., "three numbers in increasing order of magnitude", Wason, 1960) by writing examples of triplets (as well as the reason for their choice) and using the feedbacks given by the experimenter about whether or not the numbers provided by the participant conformed to the rule. There were no time constraints, but participants were asked to produce the least possible number of sets.

The results showed that only a minority of participants (6 out of 29) found out the correct rule at their first attempt. These participants tended to provide more negative examples of the rules than participants who provided a first incorrect rule. Specifically, the mean ratio of the number of negative instances to the total number of instances provided was .21 for the six participants who immediately found out the correct rule and .04 for the 22 participants who provided a first incorrect rule, $p = .0002$, one-tailed test. Most importantly, participants who gave the correct rule at their first announcement tended to use an eliminative strategy (equivalent to what Bruner et al. (1956) labeled "conservative focussing"), that is, to check for both confirming and disconfirming pieces of information so that inappropriate hypotheses about the unknown rule could be

eliminated. Indeed, the ratio of the number of inconsistent to consistent instances (i.e., the eliminative/enumerative ratio) was, on average, 1.79 for the six participants who guessed immediately and correctly the rule, and .24 for the 22 participants who provided a first incorrect rule, $p = .0002$, one-tailed test.

In other words, participants mostly adopted an enumerative strategy, wherein they generated sets of three numbers that conformed to their current working hypothesis, that is, they tended to look only for confirming evidence (i.e., Baconian enumerative induction). As Wason pointed out in the introduction and in the discussion of his article this tendency to produce "sufficient rules" is akin to what Bruner et al. (1956) labeled "successive scanning" and "direct test". Three decades later this testing procedure will be called "positive testing" (e.g., Klayman, 1995; Klayman & Ha, 1987), "congruence bias" (Baron et al., 1988), "matching strategy" (Dardenne & Leyens, 1995), "hypothesis-true questions" (Devine, Hirt, & Gehrke, 1990) (in the following section of the chapter we will describe in more details this hypothesis-testing strategy).

Indeed, it is precisely a "positive testing" what participants in Wason's (1960) experiment liked best as a procedure to discover the rule. However, the interpretation given by Wason basically rested on a fundamental confirmation bias of people who would show reluctance to adopt a falsifying strategy, which Popper (1959) claimed to be an efficient procedure to address issues in the scientific domain. Subsequent studies, spanning from Wetherick's (1962) to Klayman & Ha's (1987), made clear that this conclusion was at least an overstatement. The enumerative strategy turned out to be misleading within the specific paradigm used by Wason (1960), because the to-be-discovered rule was unrestrictive to the point that the hypotheses generated by the participants were likely to entails it, but not vice versa. In other words, the participants were likely to receive confirmations but not refutations of their hypotheses because an instance supporting their hypothesis (e.g., "three *even* numbers in increasing order of sequence", Poletiek, 2001) was also a positive instance of the rule (.e., "three numbers in increasing order of magnitude").

We will refer to the benefits that positive testing, far from being a necessarily confirmatory and maladaptive strategy, might foster under some circumstances in a later section. Now, we turn to another seminal paradigm introduced by Wason, the selection task. He conceived it during his stay at the Harvard Center for Cognitive Studies in 1963, even though it was published three years later. In a subsequent work (Wason, 1968),

Wason worked out in depth the issue outlined in the 1966's "pilot study", which was included in a chapter of the book edited by Brian M. Foss. Indeed, in "Reasoning" (1966), Wason introduced the selection task in a paragraph on errors in deductive reasoning, further deepening the alternative explanation given by Chapman & Chapman (1959) to some errors in syllogistic reasoning (specifically, the acceptance of the converse of universal affirmative, A, and particular negative, O, propositions) which, according to Chapman and Chapman, could not be fully accounted by the "atmosphere effect" advanced by Woodworth and Sells (1935) and Sells (1936), rather could be traced back to the effect of experienced procedures which are justifiable in the real world. Wason presented participants with four cards, of which they could see only one side. They were told that each card had a letter on a side a number on the other one. The task was to select the cards that would have to be turned over to falsify the statement: "if there is a vowel on one side, then there is an even number on the other side". Formally speaking, the correct answer entails choosing the card with a vowel on the front and the card with an odd number on the front. Indeed, this is the only combination which allows to falsify the rule if we conceive the task as a deductive one, wherein the rule to be tested is a conditional sentence. In other words, the rule can be expressed in logical symbols as: "if P then Q", where "P" stands for "vowel" and "Q" for "even number". When making explicit the truth table of this conditional sentence, it turns out that all of the combinations (i.e., PQ, ¬PQ, ¬P¬Q) lead to the truth of conditional, but one, that is: P¬Q. In the first version of the study and in its replication by Hughes in 1966, as reported in Wason (1968), nearly all participants selected P, from 60% to 75% of participants selected Q, a minority selected ¬Q, and almost anybody ¬P. Thus, the majority of participants committed the fallacy of affirming the consequent, that is, deducing P from Q, and did not choose the contrapositive, that is, deducing ¬P from ¬Q (a.k.a. *modus tollens*). Wason based his interpretation of these findings on two assumptions. First, participants assigned three instead of two truth values to the conditional statement, that is, "true", "false", "irrelevant" (i.e., participants had a "defective truth table", Poletiek, 2001, p. 78). This would explain the tendency to select Q (in order to assess whether it is associated to P, thus determining the truth of the conditional), and the infrequent selection of ¬P (which, in real life, is irrelevant to the truth or falsity of a conditional sentence). Furthermore, according to Wason, this assumption is consistent with the explanation given by Chapman and Chapman, because in everyday life a conditional with a false antecedent does not

appear to be true, rather a false antecedent makes irrelevant the question about the truth of the sentence (Wason, 1966). Wason made a second assumption about a tendency of participants "to expect a relation of truth, correspondence or match to hold between sentences and states of affairs" (Wason, 1968). In other words, the failure to select ¬Q (and to deduce from it ¬P) would rise from a difficulty to handle falsity or negation (e.g., Wason 1959, 1961), thus transforming what is affirmed in the rule, Q, in its negation, ¬Q (Wason, 1968).

Wason concluded that the results provided evidence that the adult participants in his task exhibited a tendency to verification (i.e., to confirm instead of eliminating hypotheses) (Wason, 1966) and did not prove to have acquired what Piaget called the "formal operational thought" (Wason, 1968).

Among the subsequent studies on the selection task, of particular interest here are those which proposed an alternative view to the logical perspective, namely a statistical-inference approach (Kirby, 1994; Oaksford & Chater, 1994). In particular, Mike Oaksford and Nick Chater (1994) advanced a rational analysis of the participants' behavior in the selection task, based on the Anderson's (1990) definition of rationality in terms of adaptation to the environment, in which the participants' choices are viewed "as optimizing the expected amount of information gained by turning each card" (Oaksford & Chater, 1994). The authors defined the expected information gain, $E(I_g)$, as the difference between the uncertainty about a hypothesis before acquiring some new data and this uncertainty after the receipt of new information. They formalized uncertainty based on Shannon's (1948) information theory, that defines uncertainty (or entropy) as:

$$E = -\sum_{i=1}^{n} p\,(\mathrm{H}_i) \log_2 p\,(\mathrm{H}_i)$$

where n is the set of alternative and mutually exclusive hypotheses and $p\,(\mathrm{H}_i)$ is the probability that each of them is the appropriate one (whenever the n hypotheses are equally probable, the formula reduces to the $\log_2(n)$). Besides assuming that participants treat the selection task as an inductive, hypothesis-testing task, Oaksford and Chater assumed that participants compare two mutually exclusive and jointly exhaustive hypotheses: The conditional sentence is true (i.e., P and Q are dependent), and the conditional sentence is false (i.e., P and Q are independent). Then they assumed that the participants consider the number of Ps and Qs in the absence of P to be constant under

whichever of the two hypotheses. Finally, the authors made a "rarity assumption", that is, they assumed that participants consider Ps and Qs in the absence of P as rare (Poletiek, 2001). Given these assumptions, Oaksford and Chater built a model of the $E(I_g)$ derived by turning over the P, Q, ¬Q, and ¬P cards in the selection cards. According to their model, the ¬P card has an $E(I_g)$ of zero, in keeping both with proposition logic and laypeople's intuitions but not with the norms in the rule discovery task, where checking ¬P might turn out to be useful to discover the unknown rule (Poletiek, 2001; Wason, 1966, 1968). The $E(I_g)$ of the P card is the highest, followed by the $E(I_g)$ of the Q card, which, in turn, is higher than that conveyed by turning over the ¬Q card when the probabilities of both P and Q are low. This ordering in $E(I_g)$ (i.e., P > Q > ¬Q > ¬P) reflects the standard frequencies of card selection. The authors demonstrated that this model captured several empirical data in the abstract as well as in the thematic versions of the selection task. Accordingly, they sanguinely conclude that their "model establishes that subjects' behavior while performing the selection task need have no negative implications for human rationality" (Oaksford & Chater, 1994).

The analysis by Oaksford and Chater has been criticized mainly because of the number and of the formality of the assumptions made, which might fail to capture the psychological mechanisms involved in the selection task (Poletiek, 2001). However, the assumption that participants "act as Bayesian optimal data selectors with rarity" (Oaksford & Chater, 1994) has fostered subsequent works which showed that some laypeople's behaviors purportedly labeled as "irrational" or "biased" might indeed be seen as a rational (Bayesian) fashion to deal with the environment (e.g., McKenzie & Mikkelsen, 2000, 2007). Not only did the Okasford and Chater's work contribute to the new perspective on rationality, but also paved the way to subsequent comparisons among different (Bayesian) optimal-experimental-design (hereafter, OED) models, of which information gain is but one (Nelson, 2005, 2008; Nelson, McKenzie, Cottrell, & Sejnowski, 2010). Much has to be done with respect to determining which normative model, and under what circumstances, best captures participants' behavior while searching for and evaluating information, but this seems a promising course of investigation to develop (as a case in point, see Crupi, Tentori, & Gonzalez, 2007).

Even though they might have led to rush conclusions about human rationality, both Wason's studies (1960, 1966) had the merit to raise the debate on whether and why people tend to preserve their own opinions (i.e., confirmation bias) and fostered a vast

range of research on how much and what kind of information people look for and make use of. Specifically, with reference to the hypothesis-testing literature, the rule discovery task addressed all the main phases of hypothesis development, that is, hypothesis generation, hypothesis testing, and hypothesis evaluation, while the selection task focused on test selection (Poletiek, 2001). We did not go through the impressive range of studies which both tasks promoted, but what is relevant here is to note that many of the experimental paradigms and concepts which have been worked out in the recent literature on hypothesis testing, if anything, have their roots in several earlier works, notably those by Bruner et al. (1956) and Wason (1960, 1966, 1968). Consider, as a specific example, the positive test strategy, which was foreshadowed first by Edna Heidbreder (1890-1985) in her PhD dissertation *An Experimental Study of Thinking* dated 1924, then by Bruner et al. (1956) who described the successive-scanning strategy, which entails direct tests, and by Wason (1960)'s enumerative test strategy[3].

In the next sections of this chapter we shall describe the main strategies that have been found to underlie people's responses in hypothesis-testing and hypothesis-evaluation tasks[4].

Testing strategies: What is a good question?

Studies on people's question-asking preferences have often used probabilistic tasks, "in which tests are formulated in probabilistic terms" (Poletiek, 2001). In this kind of tasks, search strategies are defined by manipulating the likelihoods of the LR, which are typically explicitly presented to participants along with the prior probabilities of the hypotheses to be tested. The participant's preference for a test with a certain LR allows the experimenter to determine the information-gathering strategy used by that participant (Poletiek, 2001).

Evaluating the formal usefulness of a question can be well approximated by the "feature-difference heuristic" (Nelson, 2005, footnote 2; Nelson et al., 2010; Slowiaczek et al., 1992). This strategy implies that one selects the test/query about the feature for which $|p(E \mid H) - p(E \mid \neg H)|$ is maximized and, provided that the task concerns two mutually exclusive hypotheses and two-valued features, it turns out to be tantamount to

[3] For a similar perspective on the usefulness of taking into account old heritage in theory development and as a source of ideas see Rakow, 2010, as well as the critical comments to his article by Daniel Kahneman and Jonathan Baron.

[4] Parts of these sections of the chapter are excerpts from Rusconi and McKenzie, in preparation.

the application of "impact", an OED model, regardless of the specific prior probabilities and likelihoods (Nelson, 2005, footnote 2; Nelson et al., 2010). Accordingly, it can be argued that estimating question usefulness might be easily approximated by intuitive judgment, and indeed people proved to be, if anything, sensitive to the formal diagnosticity of questions (e.g., Skov & Sherman, 1986; Slowiaczek et al., 1992; Trope & Bassok, 1982). From a formal standpoint, the usefulness of a question can be computed as the average of the likelihood ratio of the confirming answer and the likelihood ratio of the disconfirming answer, each weighted for the respective probabilities of occurrence (e.g., Nelson, 2005; Slowiaczek et al., 1992; Trope & Bassok, 1982). It is expressed as follows:

$$\{p\,(E) \times \max\,[p\,(E \mid H) / p\,(E \mid \neg H), p\,(E \mid \neg H) / p\,(E \mid H)]\} + \{p\,(\neg E) \times \max\,[p\,(\neg E \mid H) / p\,(\neg E \mid \neg H), p\,(\neg E \mid \neg H) / p\,(\neg E \mid H)]\}$$

As it will be shown in Chapter 2, some studies (e.g., Skov & Sherman, 1986; Trope & Bassok, 1982) have found that people are sensitive to the diagnosticity of questions, accordingly the label "diagnosing strategy" (Skov & Sherman, 1986; Trope & Bassok, 1982) or "diagnostic strategy" (Dardenne & Leyens, 1995; Devine et al., 1990) has been used to define people's test preferences which conform to the normative criterion. Dardenne & Leyens (1995) in a study concerning "trait hypothesis testing" (Evett et al., 1994) gave the following example of diagnostic question when testing the hypothesis of one's introversion: "Do you dislike meeting new people?" (Dardenne & Leyens, 1995). Indeed, according to the authors, such question maximizes the difference between the working hypothesis (i.e., introversion) and the competing hypothesis (i.e., extroversion).

Other two strategies have been found to play a key role in hypothesis testing. One is positive testing (e.g., Klayman, 1995; Klayman & Ha, 1987; McKenzie, 2004; Poletiek, 2001). As we have already reported above, this strategy has been identified in earlier studies on hypothesis testing under the labels of "successive-scanning strategy" or "direct test" (Bruner et al., 1956) and "enumerative thinking" (Wason, 1960). More recently, this testing strategy has been called "congruence bias" (Baron et al., 1988), "*hypothesis true* questions" (Devine et al., 1990), "matching strategy" (Dardenne & Leyens, 1995). It can be defined as a tendency to ask questions for which a "yes" answer is expected if the working hypothesis is true or, in a logically equivalent fashion, as a tendency to inquire about a feature whose occurrence is expected given the truth of the working hypothesis

(e.g., McKenzie, 2004). Formally speaking, positivity reflects a strategy wherein one select those tests about features for which p (E | H) > p (E | ¬H), where "E" is the feature under consideration, "H" is the working hypothesis, "¬H" is the alternate hypothesis (e.g., Nelson, 2008). The example provided by Dardenne & Leyens (1995) in a more contextualized study is "Do you like to stay alone?" to test the introvert hypothesis. Indeed, typically an introvert person will answer "yes" to this kind of questions.

We shall describe in Chapter 2 some inconsistencies in the literature about the extent to which people resort to this strategy. For now, let us make clear that, contrary to previous claims (e.g., Snyder & Swann, 1978; Wason, 1960), positive testing does not necessary lead to confirmation bias. Consider again the example given by Dardenne & Leyens: If one answers "yes" to that question, the tester would get confirmation of her/his working hypothesis (i.e., introversion), but a "no" answer would falsify the hypothesis (e.g., Devine et al., 1990). Indeed, instead of emphasizing the allegedly pernicious consequences of using a positive test strategy, some scholars have shed light on the bright side of positivity. For instance, Dardenne and Leyens (1995), who labeled this testing procedure "matching strategy", emphasized its confirmatory character (although they acknowledged that it "can lead to a strong disconfirmation", Dardenne & Leyens, 1995), but also that it can be both informative and useful in social interactions. Indeed, asking "matching questions" might favor "a smooth interaction" (Dardenne & Leyens, 1995). Indeed, as they pointed out, asking an introvert person if she likes reading poetry by herself (a matching question) might sound more empathic than asking if she likes crowded gigs (a nonmatching question). Fiedler and Walther (2004) summarized other aspects which make reasonable to adopt a positive test strategy. In tasks such as the 2−4−6 task described above, positive testing might be the only procedure that allows unambiguous falsifications whenever the target rule is more restrictive than the hypothesized rule (see also Klayman & Ha, 1987). Furthermore, under some circumstances positive testing allows to acquire more diagnostic information than negative testing. For example, discovering that a person is always the life of parties is more informative about her extroversion than learning that she is not (as it will be described later, this kind of questions are more properly labeled asymmetrically confirming). There are also pragmatic factors which might foster the adoption of positive testing. Sometimes it is undoable to check for all of the negative instances, because they are more numerous than the positive instances (i.e., when the working hypothesis refers to

a sample of small size). There is a third strategy which has been found to play a key role in hypothesis testing (in Chapter 2, it will be pointed out that such a prominent role is more controversial than previously thought). This has been usually referred to as "extremity" (McKenzie, 2004; Skov & Sherman, 1986; Slowiaczek et al., 1992). As we shall see in Chapter 2, extremity is equivalent to the asymmetrically disconfirming strategy. A dichotomous question is defined *symmetric* whenever the amount of information conveyed by the confirming answer (which could be either a "yes" answer or a "no" answer) is equal to that transmitted by the disconfirming answer (again, either a "yes" or a "no"). By contrast, a query is labeled *asymmetric* whenever the informativeness of the confirming answer differs from that of the disconfirming answer. Specifically, when the confirming answer weights more than the disconfirming answer the questions is called *asymmetrically confirming*. Vice versa, when the disconfirming answer conveys more information than the confirming answer the question is labeled *asymmetrically disconfirming* (e.g., Cameron & Trope, 2004; Trope & Liberman, 1996; Trope & Thompson, 1997, see Figure 1.1). Accordingly, the a/symmetric strategy is defined in relation to the informativeness of the anticipated outcomes of an inquiry.

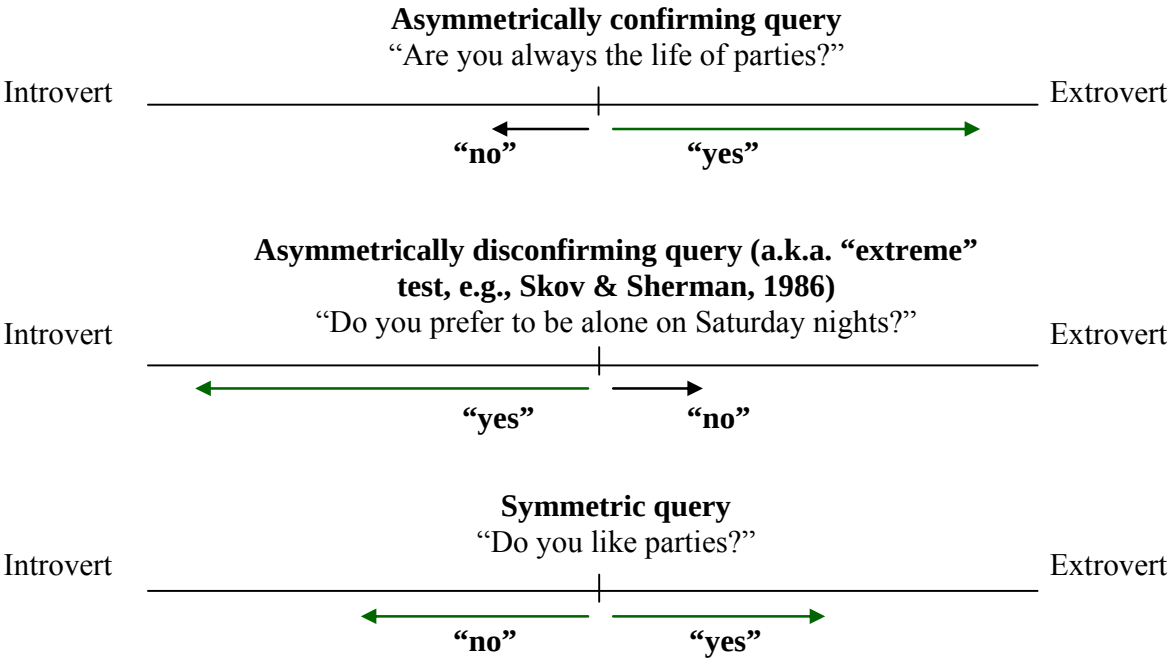


Figure 1.1. Examples of asymmetrically confirming, asymmetrically disconfirming and symmetric queries given by Trope & Thompson (1997) and graphical illustration of the "evidential value" (Poletiek & Berndsen, 2000) of the "yes" and "no" answers to such questions.

We only hint at two other testing strategies, called "*information* heuristic" and "*certainty* heuristic", introduced by Baron et al. (1988). The former refers to the tendency to deem as useful also tests which in fact are useless (Baron et al., 1988; Poletiek, 2001). The latter reflects an overestimation of queries which can confirm or eliminate one or more hypotheses with certainty (Baron et al., 1988). To many respects, Baron et al.'s (1988) study deserves attention because it took into account probabilities and hypotheses elicited from participants vs. provided by the experimenter, unequal prior probabilities, multiple hypotheses to test as well as certain tests (i.e., questions about features with either p (E | H) = 1.00 or p (E | H) = .00). However, precisely for these interesting manipulations it departs from the traditional probabilistic tasks used in the hypothesis-testing literature as well as from those we are going to illustrate in the next chapters, in which participants are given two mutually exclusive and jointly exhaustive hypotheses with equal prior probabilities to evaluate and the tests are provided in terms of likelihoods which differ from p = .00 and p = 1.00. Hence, we refer the reader to Baron et al.'s (1988) article itself and to Poletiek (2001) for further details on these strategies.

Evaluating new incoming information

Surprisingly, less attention has been devoted to hypothesis-evaluation strategies than to hypothesis-testing strategies. This is remarkable, especially if we consider that in the 90s it has been emphasized that confirmation biases are likely to occur when biases in the testing phase are coupled with biases in the evaluation stage of hypothesis development, while biases in but *one* of these two phases do not necessitate *per se* a tendency to confirmation (e.g., McKenzie, 2006; Klayman, 1995; Poletiek, 2001; Slowiaczek et al., 1992). For instance, as we have already pointed out, positive testing by itself does not necessarily lead to confirmation bias, because one can take into account the focus on an expected event under the working hypothesis (assumed to be true) once it receives an answer or when it evaluates the actual occurrence or non-occurrence of the event.

The evaluation of the informativeness of answers seems to be less easily approximated by intuitive reasoning than the estimation of question usefulness. From a Bayesian perspective (but for a critical description of other normative criteria see Nelson, 2005), the impact of the evidence is expressed through the LR, also known as Bayes factor (e.g., Nelson, 2005; Slowiaczek et al., 1992)[5]. We recall the reader that this is the

[5] Alan Turing (1912-1954) was the first to use the LR as a measure of the "evidential value" (Poletiek &

ratio of the probability of finding a piece of evidence given the truth of the working hypothesis to the probability of finding the same piece of evidence given the falsity of the working hypothesis. In other words, the evidential value of a "yes" answer (or, in a logically equivalent way, the impact of the presence of a feature) can be expressed as follows:

$$\max [p (E \mid H) / p (E \mid \neg H), p (E \mid \neg H) / p (E \mid H)]$$

while the impact of a "no" answer (i.e., the absence of a datum) is computed as:

$$\max [p (\neg E \mid H) / p (\neg E \mid \neg H), p (\neg E \mid \neg H) / p (\neg E \mid H)]$$

Consider, for example, the planetary scenario introduced by Skov and Sherman (1986, see also McKenzie, 2006; Nelson, 2008; Nelson et al., 2010; Slowiaczek et al., 1992). On an imaginary planet, Vuma, there are two kinds of inhabitants—Gloms and Fizos—which are equally numerous (i.e., the prior probability of encountering a Glom or a Fizo is .5) and invisible to human sight. The only way to identify the creatures is by asking about some features they possess. Participants are told the distribution of probabilities of the features across Gloms and Fizos. For example, participants are told that 90% of Gloms and 50% of Fizos drink gasoline. If one asks about drinking gasoline in order to determine whether the encountered creature is a Glom, the impact of a "yes" answer is .9/.5 = 1.8, while the evidential value of a "no" is .5/.1 = 5.00. Thus, the "yes" answer and the "no" answer to the same question convey a different amount of information: In particular, the "no" answer is more informative than the "yes" answer.

As we shall see in Chapter 3, the few studies that have addressed the issue of how people evaluate the informativeness of the different outcomes of the same test (McKenzie, 2006; Skov & Sherman, 1986; Slowiaczek et al., 1992) have pointed out that people tend not to revise differently their opinions in light of the different amount of information conveyed by the "yes" and "no" answers to the same question when the task has an abstract content.

Berndsen, 2000) of a datum. Specifically, he used the log likelihood ratio, which he called *weight of evidence*, whose unit of measure is *ban* (Good, 1979).

Chapter 2

Question-asking preferences in hypothesis testing[6]

The main goal of the study we shall present in this chapter is to further explore two people's tendencies in information gathering, namely the alleged preference for posing *asymmetrical questions* and *positive questions*, 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.

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. This might have important consequences. Indeed, different studies have emphasized that some human trends in gathering information might—in certain environments—cause undesirable side effects such as confirmation biases (Klayman & Ha, 1987; Nickerson, 1996, 1998; Wason, 1960, 1968) or the preservation of social stereotypes (Cameron & Trope, 2004; Trope & Thompson, 1997).

Positivity and asymmetry of questions

As shown in Chapter 1, a common definition describes a positive question as a question where a positive response ("yes") supports the truth of the hypothesis (Klayman, 1995; Klayman & Ha, 1987; Snyder & Swann, 1978). However, posing positive questions does not necessarily imply an ability to anticipate the epistemic effects of the "yes" or "no" answers. 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

[6] Most of the material presented in this chapter appears 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. This study was partly funded by a PRIN 2006 grant to Professor Paolo Cherubini by the Italian government, and by a 2007 FAR grant to Professor Paolo Cherubini 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.

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 et al., 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 et al., 2007).

The first, most common, and easiest one is the Bayes' factor, or LR (see Chapter 1). 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 $p(E \mid H) = p(E \mid \neg H)$, 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 ¬H: it is thus *confirmatory*. Finally, a piece of evidence with LR < 1 decreases the posterior probability of H with respect to ¬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 ¬H, or vice-versa). Otherwise, it is asymmetric.

Symmetric queries are "fair" questions, with equal chances[7] 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 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 pose 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

[7] Whilst the properties of LRs are independent of the priors p (H) and p (¬H), the probabilities of receiving either a confirming or a disconfirming answer depend upon p (H). Where we discuss them in this chapter we assume that p (H) = .5, a common assumption in most other previous studies on this topic, and a premise in our experiments.

hypothesis and the LR of the disconfirming response in support of the alternative hypothesis, as we have described in Chapter 1.[8]

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 sub-

[8] 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., Slowiaczek et al., 1992; Trope & Bassok, 1982). 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.

optimal *a priori*, because they can be—and often are—"optimal" in some specific environments (e.g., McDonald, 1992; Oaksford & Chater, 1994), 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. Cameron & Trope, 2004; Henrion & Fischhoff, 1986; Koehler, 1993; Lord, Ross & Lepper, 1979; Newman, Wolff, & Hearst, 1980; Poses, Bekes, Copare & Scott, 1990; Ross & Anderson, 1982; Slowiaczek et al., 1992), or the type of local environment where the evidence has been harvested (e.g., Fiedler, 2000; Fiedler, Brinkmann, & Betsch, 2000; Klayman, 1995; Klayman & Ha, 1987; Nickerson, 1998).

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.

Symmetric vs. asymmetric queries

Slowiaczek et al. (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), which we have already mentioned in Chapter 1. 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) = 1.00 and p (E | ¬H) = .00, and uninformative tests with p (E | H) = p (E | ¬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* (feature | H): asymmetrically disconfirming, negative question
2. low *p* (feature | H): almost symmetric, negative question
3. high *p* (feature | H): almost symmetric, positive question
4. very high *p* (feature | H): asymmetrically disconfirming, positive question

High Diagnosticity questions:

5. very low *p* (feature | H): almost symmetric, negative question
6. low *p* (feature | H): asymmetrically confirming, negative question
7. high *p* (feature | H): asymmetrically confirming, positive question
8. very high *p* (feature | 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 (Cameron & Trope, 2004; Trope & Thompson, 1997), it was demonstrated that when individuals try to make accurate judgments 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?").[9] 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: 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

[9] 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.

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, in which 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.

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. Poletiek and Berndsen (2000) made an attempt to reconcile those conflicting results by addressing hypotheses testing as a risk-taking behavior, 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.[10] 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

[10] 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.

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.

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. As we have already pointed out (Chapter 1), 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, 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 (Cherubini, Castelvecchio, & Cherubini, 2005; Gale & Ball, 2006; Kareev, Halberstadt, & Shafir, 1993; Klayman & Ha, 1987, 1989; Rossi, Caverni, & Girotto, 2001; Spellman, Lopez, & Smith, 1999; Vallée-Tourangeau, Austin, & Ramkin, 1995). 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 ¬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 ¬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 (see Chapter 1)—in their Experiments 1 and 2, in which the participants focused on a given hypothesis; however, the bias weakened (Experiment 3) or disappeared (Experiments 4-5) where all the alternative hypotheses where 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.

Goals of the present experiments

The present experiments empirically investigate the use of the asymmetric testing and positive testing strategies, with the aim of 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).

Experiments 1 and 2

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.

Experiment 1

Participants

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

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" judgment, similar to the dependent variable used by Skov and Sherman (1986), but different from that used by Trope and Bassok (1982). Second, 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".

Results

Table 2.1 shows the raw number of total choices (and proportions) as a function of positivity and asymmetry of the questions.

	positive questions	*negative questions*
symmetric questions	188 (39.17%)	51 (10.62%)
asymmetric questions	187 (38.96%)	54 (11.25%)

Table 2.1. Experiment 1. Total number of choices (and proportions) for each question.

There is no reliable preference for symmetric questions over asymmetric ones, or vice versa (symmetric questions: 239, asymmetric questions: 241), exact binomial test, $p = .96$. By contrast, positive questions were preferred to negative ones (positive questions: 375, negative questions: 105), $p < 0.001$. The two factors did not interact, $\chi^2 = .08$, df = 1, exact $p = .83$. 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 2.2.

	Mean Choices	*SD*
2 positive questions	4.73	3.43
2 negative questions	0.23	1.10
2 symmetric questions	1.33	1.83
2 asymmetric questions	1.37	1.65
Pos. sym. & neg. asym.	0.20	0.55
Pos. asym & neg. asym.	0.13	0.43

Table 2.2. Experiment 1. Mean number of selections (and standard deviations) for each pattern of choice, over the eight problems.

Table 2.3 shows the exact *p* values of each comparison among the six possible patterns, by means of Wilcoxon non-parametric exact tests.

	2 neg.	*2 sym.*	*2 asym.*	*Pos. sym, neg. asym.*	*Pos. asym., neg. sym.*
2 positive questions	<.001	<.001	.002	<.001	<.001
2 negative questions		.02	.01	n.s.	n.s.
2 symmetric questions			n.s.	.005	.003
2 asymmetric questions				.003	.002
Pos. sym., neg. asym.					n.s.

Table 2.3. Experiment 1. Exact *p* values for each comparison among the selections shown in Table 2.2, Wilcoxon tests.

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

	positive questions	*negative questions*
Confirming questions	98 (40.66%)	22 (9.13%)
Disconfirming questions	89 (36.93%)	32 (13.28%)

Table 2.4. Experiment 1. Distribution of the 241 asymmetric choices.

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, $p < .001$. The interaction is not significant, $\chi^2 = 2.28$, df = 1, exact 2-tailed $p = .16$.

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 $r = .53$, $N = 32$, $p < .005$, and expected Information Gain, $r = .54$, $p < .005$.

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 $r = .47$, $p < .01$, with the number of symmetric positive questions selected by each participant, and negatively, $r = -.40$, $p < .05$, with the number of negative asymmetric questions selected by each participant.

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. Second, 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 at the beginning of this chapter: Trope and his associates (Cameron & Trope, 2004; Trope & Thompson, 1997) 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.

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.

Participants

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

Results

Table 2.5 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, $p = .001$, but it also showed that symmetric questions were apparently preferred to asymmetric ones, $p = .002$.

	positive questions	*negative questions*
symmetric questions	1.58	2.28
asymmetric questions	1.75	2.37

Table 2.5. Experiment 2. Mean ranks assigned to each question. Rank "1" denoted the most useful query, while rank "4" the least useful one.

Positive symmetric questions were rated more useful than positive asymmetric, $p = .002$, negative symmetric, $p < .001$, and negative asymmetric ones, $p < .001$. Positive asymmetric questions were preferred to negative symmetric ones, $p = .03$, and negative asymmetric ones, $p = .002$. Finally, negative symmetric questions were ranked more important than negative asymmetric ones, $p = .04$. By aggregating data across participants, however, we found reliable negative correlations between the mean rank attributed to each questions and its diagnosticity, Pearson's $r = -.58$, $N = 32$, $p < .001$; 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 eight 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 2.6.

	symmetric questions	*Asymmetric questions*
Symmetric questions are more diagnostic	1.78	2.23
Symmetric questions are less diagnostic	2.17	1.78

Table 2.6. Experiment 2. Mean ranks assigned to questions in the two group of problems where symmetric questions are either more (5 problems) or less (3 problems) diagnostic than asymmetric ones. Rank "1" denoted the most useful query, while rank "4" the least useful one.

Two two-tailed Wilcoxon exact tests showed that symmetric questions were significantly preferred in the problems where they were more informative, $p < .001$, but asymmetric ones were preferred in the problems where the symmetric questions were less informative, $p = .003$. 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., Garcia-Marques, Sherman, Palma-Oliveira, 2001; Trope & Bassok, 1982), in our view it is acceptable evidence that the apparent preference for symmetric questions reported above is—at least partly—an artefact generated by the differences in the diagnosticity of the problems.

Table 2.7 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).

	positive questions	*negative questions*
Confirming questions	1.73	2.33
Disconfirming questions	1.76	2.38

Table 2.7. Experiment 2. Mean ranks assigned to asymmetric questions. Rank "1" denoted the most useful query, while rank "4" the least useful one.

Two two-tailed Wilcoxon exact tests confirmed, once again, that positive questions were preferred to negative ones, $p < .001$, 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.

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 introductory sections of this chapter). 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.

Experiment 3

Klaus Fiedler, commenting on a previous draft on these experiments, correctly noticed that, with the exception of problem 3, the numerals describing the asymmetric queries in Experiments 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.

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).

Results

Table 2.8 shows the mean ranks assigned by participants to each question (averaged over the eight problems).

	positive questions	*negative questions*
symmetric questions	1.72	2.41
asymmetric questions	1.80	2.46

Table 2.8. Experiment 3. Mean ranks assigned to each question. Rank "1" denoted the most useful query, while rank "4" the least useful one.

A set of two-tailed Wilcoxon exact tests confirmed that positive questions were preferred to negative questions, $p < .001$, and that symmetric questions were apparently preferred to asymmetric ones, $p = .03$, fully replicating the main findings of Experiment 2. Positive symmetric questions were rated marginally more useful than positive asymmetric ones, $p = .06$, and significantly more useful than negative symmetric and asymmetric ones, $p < .001$ for both comparisons. Positive asymmetric questions were preferred to negative symmetric and asymmetric ones, $p < .001$. 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, $r = -.55$, $p < .001$; 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 were 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, $p = .003$ for the five problems where symmetric questions are more informative, and $p < .001$ where asymmetric questions are the more informative ones. Mean ratings are shown in Table 2.9.

	symmetric questions	*Asymmetric questions*
Symmetric questions are more diagnostic	1.96	2.23
Symmetric questions are less diagnostic	2.24	1.97

Table 2.9. Experiment 3. Mean ranks assigned to questions in the two group of problems where symmetric questions are either more (5 problems) or less (3 problems) diagnostic than asymmetric ones. Rank "1" denoted the most useful query, while rank "4" the least useful one.

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 2.10 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).

	positive questions	*negative questions*
Confirming questions	1.81	2.43
Disconfirming questions	1.80	2.49

Table 2.10. Experiment 3. Mean ranks assigned to asymmetric questions. Rank "1" denoted the most useful query, while rank "4" the least useful one.

Two two-tailed Wilcoxon exact tests confirmed that positive questions were preferred to negative ones, $p < .001$, 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.

Discussion

In Experiments 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.

Experiment 4

In the three previous experiments, two negative findings were recurrently observed, suggesting that they are worth of notice. First, 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 introductory sections of this chapter). 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 (Cameron & Trope, 2004; Trope & Thompson, 1997) or for disconfirmatory queries (Skov & Sherman, 1986; Slowiaczek et al., 1992). This was not the case in the Experiments 1-3. In aggregated form, participants' choices were indifferent to the confirming or disconfirming valence of the asymmetric questions. The designs of Experiments 1-3 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.

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.

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 ($p > .9$), high ($.5 < p < .9$), low ($.5 > p > .1$), and very low ($p < .1$). 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 .68 of being present in Deck A, and of .18 in Deck B, and the participant is focusing on Deck A, the most extreme probability (.18) is in the "low" range, and is associated to the non-focal hypothesis. As in another example, if p (square | Deck A) = .98, and p (square | Deck B) = .89, the extreme probability (.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 × type of asymmetry

design (see Appendix G). Furthermore, by varying probabilities over four magnitudes (instead of two: i.e., $p > .5$ vs. $p < .5$), 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. .87). However, the diagnosticity of questions within each problem was kept constant.

Results

Results are shown in Table 2.11.

	positive questions	*negative questions*
Confirming questions	1.89	2.12
Disconfirming questions	1.65	2.32

Table 2.11. Experiment 4. Mean ranks assigned to questions (all questions were asymmetric). Rank "1" denoted the most useful query, while rank "4" the least useful one.

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 $>_{p = .011}$ positive confirming $>_{p = .014}$ negative confirming $>_{p = .034}$ negative disconfirming (all of the remaining pair-wise comparisons among 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, that is, 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.[11] Namely, the most preferred positive disconfirming questions had, on average, a *p* (E | H) greater than the second-rated positive 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 2.12.

	Extreme probability is:			
	Very low	*Low*	*High*	*Very high*
Extreme probability is associated to the:				
Focal hypothesis	2.41	2.31	1.64	1.62
Non-focal hypothesis	1.99	1.78	2.14	1.99

Table 2.12. Experiment 4. Mean ranks assigned to questions as a function of whether the extreme probability was *p* (E | H) or *p* (E | ¬H), and of its magnitude. Rank "1" denoted the most useful query, while rank "4" the least useful one.

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), $p = .001$ and $p < .001$, 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, $p = .044$. 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), $p < .001$ and $p = .005$, 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

[11] 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).

very low extreme questions, $p = .003$). 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's $r = -.71$, $p < .001$, whereas p (E | ¬H) did not, $r = -.18$, 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 | ¬H), obtaining positive values for positive questions and negative values for negative ones] significantly fits the data, $r = .77$, $r^2 = .59$, $p < .001$, with the following parameters:

mean rank = 2.28 − .57 p (E | H) − .42 [p (E | H) − p (E | ¬H)]

By adding to the model, as predictors, p (E | ¬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.

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 | ¬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 | ¬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).

General discussion

The study of spontaneous hypothesis-testing strategies has a long history in cognitive psychology, dating back to its early years (e.g., Bruner et al., 1956; Wason, 1960, see Chapter 1). 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., McKenzie, 2004; Nickerson, 1998). 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. Second, 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., Baron et al., 1988; Poletiek & Berndsen, 2000; Slowiaczek et al., 1992), 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 Baron et al., 1988; Trope & Bassok, 1982), 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 (Cameron & Trope, 2004; Poletiek & Berndsen, 2000; Trope & Thompson, 1997), 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 judgments. 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 1-3 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 Experiments 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 the analyses on asymmetric queries only, we did not find any systematic preference either for confirming, or for disconfirming queries. In Experiments 1-3, 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—within 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 question. 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 behavior reminiscent of "pseudodiagnostic" judgments (e.g., Doherty & Mynatt, 1986; Doherty, Mynatt, Tweney, & Schiavo, 1979), 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.

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; Cameron & Trope, 2004) 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 1-3—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 behavior in unfamiliar, abstract contexts is driven by simple, intuitive evaluations, more than complex, analytical ones.

In conclusion, in 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.

In this chapter, we have examined the hypothesis-testing phase of hypothesis development. In the next chapter, we shall turn to how people behave in the subsequent stage, namely hypothesis evaluation. Specifically, we shall consider how people revise

their initial opinions in light of answers to dichotomous questions (i.e., which admit only "yes" or "no" answers).

Chapter 3

Insensitivity and oversensitivity to answer diagnosticity[12]

As we have already shown (see Chapter 1), hypothesis development is a multi-componential process (e.g., Evett et al., 1994; Klayman, 1995; McKenzie, 2004), that encompasses how people generate hypotheses, seek out evidence and evaluate the available information in order to confirm or revise their pre-existing expectancies.

In Chapter 2, we have addressed the issue of how people choose which test is most useful to decide which of two mutually exclusive and jointly exhaustive hypotheses is the most appropriate. Indeed, most of the literature on hypothesis development has dealt with the testing phase and shed light on the strategies used by people when determining which is the most useful question to ask (e.g., Baron et al., 1988; Devine et al., 1990; Skov & Sherman, 1986; Snyder & Swann, 1978; Trope & Bassok, 1982). This is a crucial step of hypothesis development because, whether deliberately or not, people can determine what kind and how much information to acquire. For example, people might seek out certain pieces of information while overlooking others. This might foster hypothesis preservation, which, in some circumstances, has pernicious consequences in terms of confirmation of fallacious or maladaptive beliefs. Indeed, although there is no question-asking strategy that *per se* can lead to overconfidence in an upheld hypothesis (McKenzie, 2006), as we pointed out in Chapter 1, certain combinations of testing and evaluation strategies do (Klayman, 1995; McKenzie, 2004, 2006; Slowiaczek et al., 1992; Zuckerman et al., 1995). Accordingly, even an optimal selection of tests or questions to ask does not guarantee an optimal belief revision/confirmation, because one can expose oneself to all the available pieces of information, but then misweigh the gathered evidence in the evaluation phase (Slowiaczek et al., 1992).

The two experiments we are going to illustrate in this chapter aimed to deepen the investigation of how people treat the acquired information. Specifically, we examined how people revise their confidence in a given hypothesis in light of different answers to the same question in an abstract task. We compared different strategies which might

[12] Most of this chapter is made up of materials which appear in Rusconi, P., & McKenzie, C. R. M. (in preparation). Testing different accounts of insensitivity and oversensitivity to differentially informative answers in abstract hypothesis testing.

underlie the relative insensitivity to differentially diagnostic answers emerged in previous studies with abstract material (McKenzie, 2006; Skov & Sherman, 1986; Slowiaczek et al., 1992).

Evidence of insensitivity to differentially diagnostic answers

As noted in Chapter 1, there are only a few studies which have directly investigated how people revise their beliefs in light of different answers to the same question (McKenzie, 2006; Skov & Sherman, 1986; Slowiaczek et al., 1992). Overall, they showed that people appreciate that a "yes" and a "no" answer convey different amount of information but they underestimate this difference in tasks with abstract materials. The first evidence of such insensitivity came from a study outlined by Skov and Sherman (1986) in the discussion of their seminal work (p. 118). Only 43% of participants showed the asymmetry of confidence in the normatively expected direction after a "yes" answer and after a "no" answer, but many were not asymmetric enough. Slowiaczek and co-workers (1992) worked out the issue in depth and found that, on average, participants estimated a difference of 6% between the probability judgments after a "yes" and after a "no", while the normative difference was of 19% (Experiment 1A, Slowiaczek et al., 1992). McKenzie (2006) replicated the findings of the study by Slowiaczek et al. (1992), but only with abstract materials (i.e., planetary scenarios). When participants were presented with familiar materials (i.e., scenarios about male and female heights) the extent of insensitivity to differentially diagnostic answers was less marked.

Although the familiarity of the materials used in the experiments turned out to be an important moderator of people's sensitivity to the differential diagnosticity of answers, it remains unclear how people behave in tasks with abstract materials. Skov and Sherman (1986) hinted at a possible relation between people's failure to perceive the asymmetry in informativeness of different answers and the failure to consider base rates, but they did not develop this idea and they did not test it empirically[13]. Slowiaczek et al. (1992) advanced an explanation basically based on the confusion of the assessment of answer diagnosticity with the assessment of question usefulness, which, according to the authors, might be related to the use of the heuristic of representativeness (e.g., Kahneman &

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

Tversky, 1972; Tversky & Kahneman, 1974). In their words: "People underestimate differences in the diagnosticity of "yes" and "no" answers to the same question, because the difference in percentages is the same for both answers" (Slowiaczek et al., 1992, p. 394).

Overview of the experiments

The aim of the two experiments was threefold. First, we wanted to further specify the investigation of previous studies on people's relative insensitivity to differentially diagnostic answers by comparing the predictions drawn by the use of different strategies that participants might use when evaluating answers in abstract hypothesis testing. Indeed, previous studies have never empirically investigated and pitted against one another possible accounts of insensitivity to answer diagnosticity.

Second, we examined how people revise their initial beliefs in light of different answers to the same question by taking into account the cases in which the two answers (i.e., "yes" and "no") convey the same amount of information. Indeed, previous studies focused on the differential impact of answers to asymmetric questions, for which a "yes" is more informative than a "no" or vice versa (for details on asymmetry of questions see Cameron & Trope, 2004; Cherubini et al., 2010; Trope & Liberman, 1996; Trope & Thompson, 1997)[14]. However, it has not been empirically tested yet whether people show the normatively expected symmetry of confidence after a "yes" answer and a "no" answer to a symmetric question. Indeed, only following symmetric queries both the probability of occurrence and the informativeness are identical for "yes" and "no" answers (e.g., Cherubini et al., 2010). In other words, symmetric questions do not imply the trade-off between probability of occurrence and diagnosticity that the other types of questions do (McKenzie, 2006; Poletiek, 2001, chaps. 1 and 2; Poletiek & Berndsen, 2000). This issue is relevant because it might clarify whether people's relative insensitivity to the differential diagnosticity of answers indicates only a tendency to perceive different answers as equally diagnostic (i.e., *underestimation* of differential diagnosticity), or also as a failure to appreciate when different answers convey the same amount of information

[14] Skov and Sherman (1986) tested the asymmetry of confidence after a "yes" and after a "no" to a question about a feature with the 90-50 percentage combination and to a question about a feature with the 10-50 percentage combination. Slowiaczek et al. (1992) used the following percentage combinations: 10-50, 90-50, 50-90, 50-10. McKenzie (2006) used the 50-2 and the 10-.1 percentage combinations in Experiment 1, the 90-99.9 and the .1-10 percentage combinations in Experiment 2.

(i.e., *overestimation* of different evidential strength), thus representing a more general failure in information use.

Finally, we aimed to investigate whether the presentation format of the distributions of probabilities of the features inquired about could affect participants' performance. For reasons we discuss later, we hypothesized that presenting the distributions of probabilities related to both the presence and the absence of the feature (i.e., p (E | H) and p (¬E | H), where "H" stands either for the working hypothesis or for the competing hypothesis), instead of only the probabilities about the presence of the feature (i.e., p (E | H)), should reduce the relative insensitivity to the differential diagnosticity of answers.

Accordingly, we set up two experiments sharing the same design, materials, procedure, and instructions. The only difference between the two experiments was in the presentation format of the distribution of probabilities of the features inquired about.

Experiment 1

Participants

One-hundred and ten undergraduate students at the University of California, San Diego (73 female, 37 male, mean age = 20.1 years, range: 18-28 years; 82 were native speakers of English) took part in the study in exchange for course credit.

Materials and procedure

We set up a planetary scenario similar to the one originally introduced by Skov and Sherman (1986) and thereafter widely used in the literature on hypothesis testing (e.g., McKenzie, 2006; Nelson, 2005; Nelson, 2008; Nelson et al., 2010; Slowiaczek et al., 1992; Villejoubert & Mandel, 2002; see also the example given in the final section of Chapter 1). Specifically, we asked participants to imagine to travel to a planet, Vuma, where there is an equal number of two kinds of creatures, called Gloms and Fizos. Participants are presented with the answers to some questions about a series of features which Gloms and Fizos possess with different probabilities. The task is to surmise whether an encountered creature is a Glom (or a Fizo, according to the version of the questionnaire) based on the priors (50% of the creatures on the planet are Gloms, 50% are Fizos), the distributions of probabilities of the features inquired about, and the answers (i.e., "yes" or "no") to the questions asked about these features (a sample stimulus is given in Appendix H).

We devised a 4 X 2 X 2 X 2 design. Within-participants variables were the type of test (test 1 about a feature with probabilities of .65 and .35 under the two hypotheses, respectively; test 2: .85-.15; test 3: .98-.5; test 4: .5-.02; see Table 3.1) and the answer that participants received ("yes" vs. "no"), whereas between-participants factors were the test order (test 1-"yes", test 2-"no", test 3-"yes", test 4-"no", test 1-"no", test 2-"yes", test 3-"no", test 4-"yes", and the completely reversed order) and the focal hypothesis (Glom vs. Fizo). Accordingly, there were four parallel versions of the questionnaire and each participant responded to eight problems, each presented in a separate page of the booklet.

	Focal hypothesis	p (E | H_1) and p (E | H_2)	Answer	Diagnosticity (LR)	Hypothesis supported	Normative confidence in the supported hypothesis
test #1	Gloms	.65	*yes*	1.86	Glom	.65
	Fizos	.35				
test #1	Gloms	.65	*no*	1.86	Fizo	.65
	Fizos	.35				
test #2	Gloms	.85	*yes*	5.67	Glom	.85
	Fizos	.15				
test #2	Gloms	.85	*no*	5.67	Fizo	.85
	Fizos	.15				
test #3	Gloms	.98	*yes*	1.96	Glom	.66
	Fizos	.5				
test #3	Gloms	.98	*no*	25.00	Fizo	.96
	Fizos	.5				
test #4	Gloms	.5	*yes*	25.00	Glom	.96
	Fizos	.02				
test #4	Gloms	.5	*no*	1.96	Fizo	.66
	Fizos	.02				

Table 3.1. The structure of the problems used in the two experiments. Answer diagnosticity is computed as likelihood ratio (Nelson, 2005). Participants in Experiment 1 were presented only with p (E | H_1) and p (E | H_2), while in Experiment 2 they received both p (E | H_1), p (E | H_2) and their complements p ($\neg$E | H_1) and p ($\neg$E | H_2).

The tests were chosen so that there were two questions (i.e., tests 1 and 2) for which the "yes" answer was exactly as informative as the "no" answer. As can be seen in Table 3.1, tests 1 and 2 differ in the evidential strength of the single answers to them. Indeed, the "yes" and "no" answers to test 1 (65-35 percentage combination) have a LR of 1.86, while the "yes" and "no" answers to test 2 (85-15 percentage combination) have a LR of 5.67. The only other difference between tests 1 and 2 is in their diagnosticity. Tests 1 and

2 do not allow us to assess whether participants underestimate the difference in diagnosticity between different answers (because the normatively expected difference is zero), but only if they are calibrated or they overestimate this null difference. Accordingly, and for the sake of comparison with the literature, we presented participants also with "yes" and "no" answers to asymmetric questions. Tests 3 and 4 are equally informative (question's diagnosticity is equal to 7.95 for both), and, for both tests, the normatively expected difference in confidence after a "yes" and after a "no" is of .3 (.96 minus .66, see Table 3.1). However, the asymmetry of confidence is not in the same direction in the two tests: In test 3 the "no" answer weights more (LR = 25.00) than the "yes" answer (LR = 1.96) and vice versa in test 4 (LR_{yes} = 25.00; LR_{no} = 1.96). In other words, test 3 (98-50 feature) is an asymmetrically disconfirming question (a.k.a. "extreme" test, Skov & Sherman, 1986; Slowiaczek et al., 1992, or "low-risk test", Poletiek & Berndsen, 2000), while test 4 (50-2 feature) is an asymmetrically confirming question ("high-risk test", Poletiek & Berndsen, 2000).

Results and discussion

A/symmetry of confidence after a "yes" and after a "no". Following the procedure used by Slowiaczek et al. (1992) and McKenzie (2006) we recoded participants' estimates with respect to the hypothesis favored by the answer. Hence, for example, a participant might receive a "no" answer to test 4 (about the 50-2 feature) under the "Glom" focal hypothesis and she/he might provide an estimate of 30% chance that the encountered creature is a Glom. This would be recoded as a 70% chance of encountering a Fizo.

On this dependent variable, for each test, we performed a series of paired *t*-tests which showed that for test 1 (about the 65-35 feature) the difference in the estimates after a "yes" (M = 59.71, SD = 17.5) and after a "no" (M = 54.18, SD = 19.4) was marginally significant, $t(108) = 1.78$, $p = .078$, $d = .3$ (95% CI [-.62, 11.7]) (Figure 3.1, top left panel). Similarly, for test 2 (85-15 feature), the difference in confidence after a "yes" (M = 72.05, SD = 22.56) and after a "no" (M = 63.87, SD = 28.14) was significant, $t(109) = 2.19$, $p = .031$, $d = .32$ (95% CI [.77, 15.6]) (Figure 3.1, top right panel). Hence, when evaluating answers to symmetric questions participants tended to perceive a difference between the informativeness of the "yes" answer and that of the "no" answer that is not normatively grounded, since the normative criterion predicts a null difference in the estimates after a "yes" and after a "no" to such questions.

For the two asymmetric tests (i.e., tests 3 and 4) the normative difference is 30. A paired *t*-test revealed that for the asymmetrically disconfirming question (98-50 feature) there was not a significant difference between the mean estimated confidence after a "yes" (M = 66.14, SD = 20.38) and that after a "no" (M = 60.56, SD = 32.09), $t(109)$ = 1.54, p = .13 (95% CI [-1.6, 12.75]) (Figure 3.1, bottom left panel). As shown by a one-sample *t*-test, the mean estimated difference (i.e., the dependent variable was the mean of: estimate after "yes" - estimate after "no") of 5.58 (SD = 37.96) was significantly less than the normative difference of 30, $t(109)$ = -6.75, p = .001, d = .64 (95% CI [-31.6, -17.25]). By contrast, participants appreciated that the "yes" and the "no" answers conveyed a different amount of information when they received the answers to the asymmetrically confirming question (50-2 percentage combination). Indeed, a paired *t*-test showed that there was a significant difference between the mean estimates after a "yes" (M = 66.13, SD = 32.29) and after a "no" (M = 55.73, SD = 28.77), $t(109)$ = 2.17, p = .032, d = .34 (95% CI [.91, 19.89]) (Figure 3.1, bottom right panel). However, as shown by a one-sample *t*-test, the mean difference of 10.4 (SD = 50.2) was significantly less than that normatively expected of 30, $t(109)$ = -4.1, p = .001, d = .39 (95% CI [-29.09, -10.11]).

Hence, overall, participants perceived a difference in informativeness between the "yes" and the "no" answers to the same (symmetric) question (i.e., tests 1 and 2) when actually the two answers were equally diagnostic. By contrast, when revising their confidence in light of answers to asymmetric questions (i.e., tests 3 and 4) they either failed to appreciate the differential diagnosticity (when the question was asymmetrically disconfirming: test 3, 98-50 feature) or they provided estimates which showed the asymmetry in confidence in the normatively expected direction, but insufficiently (when the query was asymmetrically confirming: test 4, 50-2 feature).

Both oversensitivity to answer diagnosticity, when the answers came from a symmetric query, and the different magnitude of insensitivity found between the evaluation of answers to the asymmetrically confirming vs. disconfirming questions cannot be accounted for by an interpretation based on the confusion of the assessment of answer diagnosticity with the assessment of question usefulness. Indeed, this explanation predicts the same pattern of responses for the two asymmetric questions and a tendency to calibration for symmetric questions.

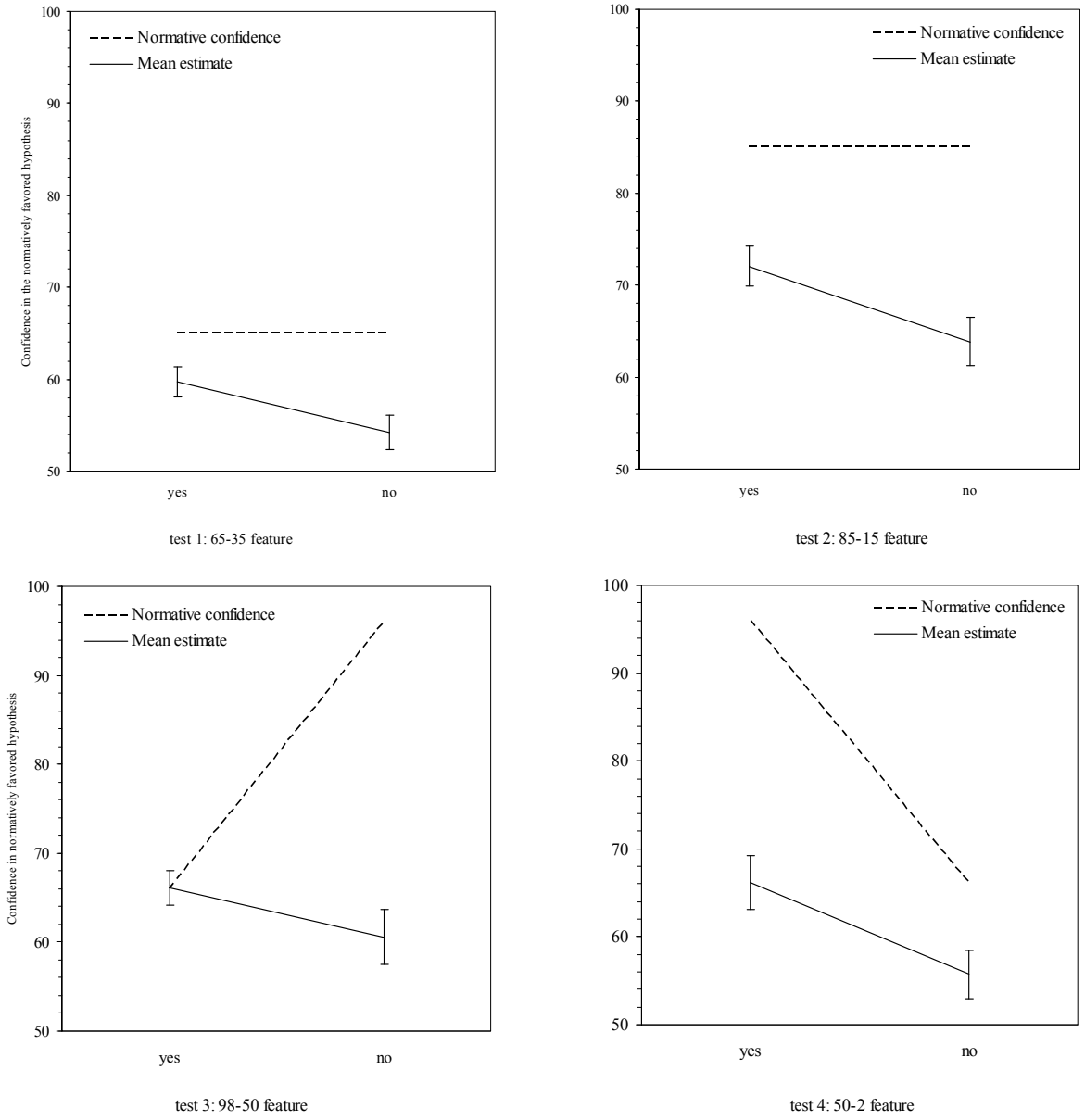


Figure 3.1. Confidence in the hypothesis favored by the evidence. The participants' mean estimates are compared with the normatively expected confidence. For participants' mean estimates, standard error of the mean (SEM) bars are also shown. Experiment 1.

Comparing predictors of insensitivity/oversensitivity to answer diagnosticity. Table 3.2 summarizes the four strategies and their predictions about confidence levels after a "yes" and after a "no", which we compared.

Strategies	Answer	Predictions	test #1: 65%-35%	test #2: 85%-15%	test #3: 98%-50%	test #4: 50%-2%
Matching heuristic	*yes*	confusion of p (H | E) with p (E | H)	65	85	98	50
	no	confusion of p(H | ¬E) with p (E | H)	35	15	50	2
Average between the likelihood and the posterior probability	*yes*	p (E | H) ≤ estimate ≤ p (H | E)	[65, 65]	[85, 85]	[98, 66]	[50, 96]
	no	p (E | H) ≤ estimate ≤ p (H | ¬E)*	[35, 65]	[15, 85]	[50, 96]	[2, 66]
Average between the prior and the posterior probabilities	*yes*	p (H) ≤ estimate ≤ p (H | E)	[50, 65]	[50, 85]	[50, 66]	[50, 96]
	no	p (H) ≤ estimate ≤ p (H | ¬E)	[50, 65]	[50, 85]	[50, 96]	[50, 66]
Average between the prior probability and the likelihood	*yes*	p (H) ≤ estimate ≤ p (E | H)	[50, 65]	[50, 85]	[50, 98]	[50, 50]
	no	p (H) ≤ estimate ≤ p (¬E | H)	[50, 65]	[50, 85]	[50, 50]	[50, 98]

Table 3.2. The strategies (and their predictions) pitted against one another in Experiment 1. For each test that we used point values or intervals resulting from each account's predictions are provided. "H" stands for the normatively favored hypothesis (i.e., "Glom" when the answer was "yes", "Fizo" when the answer was "no"). Note that the normative likelihood for the "no" answer is p (¬E | H). However, in Experiment 1, we considered the average between the likelihood and the posterior probability as a strategy wherein people average the likelihood they received in the scenario (i.e., p (E | H)) and the normative posterior probability.

We decided to consider these strategies because they predict underestimation or overestimation of the normative asymmetry of confidence when applied to the distributions of probabilities used in previous studies on sensitivity to answer diagnosticity (see Table 3.3)[15]. Specifically, the matching heuristic—originally introduced by Evans (1972) in the domain of propositional reasoning—has been

[15] For the averaging strategies that do not lead to point predictions but to intervals of confidence, the difference between the informativeness of the "yes" answer and the informativeness of the "no" answer was computed as the absolute value of the difference between the midpoint of the interval predicted for the "yes" answer and the midpoint of the interval predicted for the "no" answer.

suggested to play a role in belief updating (Villejoubert & Mandel, 2002; see also Chapter 2)[16]. It predicts that people would consider as relevant the information which matches the rule to be tested (in our task the probabilities of the features in the two groups). The other three accounts are strategies based on averages: The average between the likelihood and the posterior probability, the average between the prior and the posterior probabilities, and the average between the prior probability and the likelihood.

Reference		Difference in confidence after a "yes" answer and a "no" answer				
	Percentage combinations	Normative	Matching heuristic	Average between the likelihood and the posterior probability	Average between the prior and the posterior probabilities	Average between the prior probability and the likelihood
Skov & Sherman (1986), Slowiaczek et al. (1992)	90-50	19	40	10.5	9.5	20
	10-50	19	40	29.5	9.5	20
Slowiaczek et al. (1992)	50-90	19	40	10.5	9.5	20
	50-10	19	40	29.5	9.5	20
McKenzie (2006)	50-2	30	48	39	15	24
	10-.1	46	9.9	27.95	23	44.95
	90-99.9	46	9.9	18.05	23	44.95
	.1-10	46	9.9	27.95	23	44.95

Table 3.3. Differences in confidence after a "yes" answer and a "no" answer according to the normative (Bayesian) criterion, and four other strategies, computed on the basis of the distributions of probabilities used in literature.

Previous research has shown that in Bayesian inference tasks people might recruit averaging strategies, which have also proved to perform accurately (McKenzie, 1994). In particular, it has been suggested that people might average the prior probabilities with the likelihoods, either when participants are provided with base rates and likelihoods in order to make a single judgment or when they are required multiple judgments in light of different pieces of information (McKenzie, 1994). Furthermore, averaging leads to "conservatism" (e.g., Edwards, 1968), that is to judgments which are closer to the chance level (p = .5) than normatively expected (e.g., McKenzie, 1994, footnote 6). In our experiment, we found evidence of conservatism, especially so when the incoming

[16] Contrary to Villejoubert and Mandel (2002), who used a planetary scenario as ours to test people's strategies when estimating posterior probabilities, we did not consider the inverse fallacy because it never predicts insensitivity but only oversensitivity or calibration (in the case of symmetric tests) when applied to our tests or to those used in previous studies.

evidence was highly diagnostic (e.g., the "no" answer to test 3 and the "yes" answer to test 4)[17]. Hence, averaging strategies seem to be appropriate candidates to account for our data. The three averaging accounts do not lead to point predictions as matching and the normative criterion do, but to intervals of confidence (see Table 3.2). Accordingly, in order to pit the strategies against one another, for each predicted value we derived intervals of confidence of comparable size, that is ±5 points on the 0-100 scale with respect to the point prediction. Consider, for instance, the prediction of the matching heuristic for the "yes" answer to test 4 (50-2 feature): The normatively favored hypothesis is "Glom", thus matching leads to a chance of 50 out of 100, because it predicts that people should cling to the likelihoods they were given. We thus considered a participant's estimate falling within the interval of confidence of [45, 55] to be consistent with the use of this strategy. In some cases the intervals were narrower than 10 points due to the limits of the scale (e.g., for the "yes" answer to test 3 about the 98-50 feature the matching heuristic predicts a point value of 98, thus the interval was $93 < x \leq 100$). In some other cases the intervals were larger than 10 points because the intervals predicted by some of the averaging accounts were larger than 10 points (e.g., "no" answer to test 1, see Table 3.4). In order to determine whether and which strategy could best account for the participants' insensitivity/oversensitivity to answer diagnosticity, we computed for each participant the mean proportion of hits, that is the mean proportion of the number of times in which the participant's estimates fell within each of the intervals of confidence predicted by each account. Based on this, we then categorized each participant in terms of the strategy that best and worst accounted for her/his responses. Specifically, for each participant, we considered the highest mean proportion of hits across accounts as indicative of the best strategy for that participant. Conversely, the lowest mean proportion of hits was indicative of the strategy that worst accounted for the participant's estimates. We admitted ties across strategies. Consider, for example, a participant whose mean proportions of hits were of .5 with respect to the normative criterion, .5 for the matching heuristic, .875 for the average between the likelihood and the posterior probability, .750 for the average between the prior and the posterior probabilities, and .875 for the average between the prior probability and the likelihood. The strategies that best account for the responses of this participant would be the average between the likelihood and the

[17] The finding of an increased conservatism for increased levels of diagnosticity of the observations has been shown in several previous studies on Bayesian inference (e.g., Edwards, 1968; McKenzie, 2006; Slowiaczek et al., 1992, Experiments 1A and 1B).

posterior probability and the average between the prior probability and the likelihood. On the contrary, the strategies that would be considered the worst predictors of the participant's performance would be the matching heuristic and the Bayesian account.

Based on this categorization of the participants, we then examined the percentage of participants for whom each strategy was the best/worst account of their performance (for a similar procedure see Villejoubert & Mandel, 2002, Figure 3). In order to determine which strategy best captured the participants' responses we considered the ratio of the percentage of participants for whom the strategy represented the best account to the percentage of participants for whom the same strategy was the worst account. Of course, the greater is the ratio, the best the strategy accounts for the responses. Furthermore, a ratio < 1 means that the strategy under consideration is the worst account for more participants than those for whom it is the best strategy.

Strategies	Answer	Predictions	test #1: 65%-35%	test #2: 85%-15%	test #3: 98%-50%	test #4: 50%-2%
Normative	*yes*		$60 < x < 70$	$80 < x < 90$	$61 < x < 71$	$91 < x \leq 100$
	no		$50 < x < 80$	$80 < x < 90$	$91 < x \leq 100$	$61 < x < 71$
Matching heuristic	*yes*	confusion of p (H | D) with p (D | H)	$60 < x < 70$	$80 < x < 90$	$93 < x \leq 100$	$45 < x < 55$
	no	confusion of p(H | ¬D) with p (D | H)	$20 < x < 50$	$10 < x < 20$	$45 < x < 55$	$0 \leq x < 7$
Average between the likelihood and the posterior probability	*yes*	p (D | H) ≤ estimate ≤ p (H | D)	$60 < x < 70$	$80 < x < 90$	$66 \leq x \leq 98$	$50 \leq x \leq 96$
	no	p (D | H) ≤ estimate ≤ p (H | ¬D)	$35 \leq x \leq 65$	$15 \leq x \leq 85$	$50 \leq x \leq 96$	$2 \leq x \leq 66$
Average between the prior and the posterior probabilities	*yes*	p (H) ≤ estimate ≤ p (H | D)	$50 \leq x \leq 65$	$50 \leq x \leq 85$	$50 \leq x \leq 66$	$50 \leq x \leq 96$
	no	p (H) ≤ estimate ≤ p (H | ¬D)	$50 \leq x \leq 65$	$50 \leq x \leq 85$	$50 \leq x \leq 96$	$50 \leq x \leq 66$
Average between the prior probability and the likelihood	*yes*	p (H) ≤ estimate ≤ p (D | H)	$50 \leq x \leq 65$	$50 \leq x \leq 85$	$50 \leq x \leq 98$	$45 < x < 55$
	no	p (H) ≤ estimate ≤ p (¬D | H)	$50 \leq x \leq 65$	$50 \leq x \leq 85$	$45 < x < 55$	$50 \leq x \leq 98$

Table 3.4. The intervals of confidence pitted against one another and with participants' estimates in Experiment 1.

The results showed that, for test 1 (65-35 feature), the average between the likelihood and the posterior probability turned out to be the strategy with the highest best/worst ratio (1.31), followed by the other two averaging strategies (both 1.16), the average between the prior and the posterior probabilities and the average between the prior probability and the likelihood, which lead to the same predictions for symmetric tests, being the likelihoods and the posterior probabilities equal when the questions are symmetric. The Bayesian strategy was the worst account for more participants than those for whom it was the best account (ratio: .91). The worst account turned out to be the matching heuristic (ratio: .28), indeed for 78.18% of participants this represented the strategy that less captured their responses (Figure 3.2, top left panel). A similar, even more clear-cut, pattern emerged when considering the more diagnostic symmetric question, that is test 2 (85-15 feature): The best account was the average between the likelihood and the posterior probability (ratio: 3.43), followed by the other two averaging strategies (both 1.50). The Bayesian strategy had a ratio of .63 (for 59.09% of participants it represented the worst account for their responses), while the worst predictor was the matching heuristic (.1), with 90.91% of participants for whom the latter strategy was the worst account (Figure 3.2, top right panel).

The averaging strategies turned out to best captured participants' responses also when taking into consideration the asymmetric questions. However, compared with the symmetric tests, it was no more the average between the likelihood and the posterior probability, rather the average between the prior probability and the likelihood which best accounted for the data. Specifically, as to test 3 (98-50 feature), the strategy that averages the prior probability with the likelihood had the highest ratio (3.25), closely followed by the average between the likelihood and the posterior probability (2.61). The other strategies had a ratio < 1. In particular, the average between the prior and the posterior probabilities had a ratio of .62, the Bayesian strategy had a ratio of .54, and, again, the worst strategy was the matching heuristic with a ratio of .37 (Figure 3.2, bottom left panel). A similar pattern emerged when examining test 4 (50-2 feature), with the matching heuristic as worst predictor (ratio: .22), and the Bayesian account and the average between the prior and the posterior probabilities with ratios < 1 as well (Bayesian: .75; average between the prior and the posterior probabilities: .58). The best account turned out to be the average between the prior probability and the likelihood

(1.86), followed by the average between the likelihood and the posterior probability (1.76) (Figure 3.2, bottom right panel).

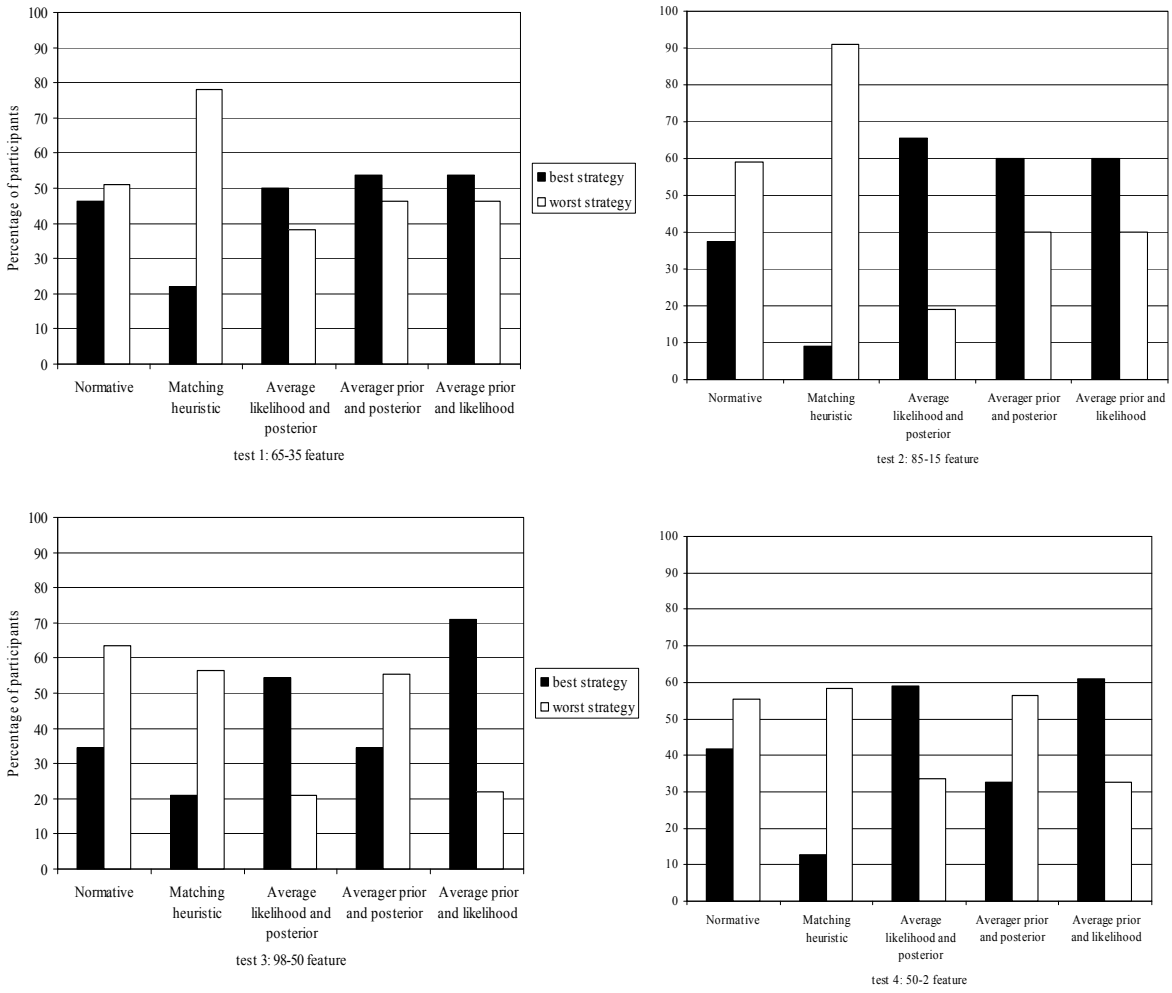


Figure 3.2. Distributions of participants, for each test, as a function of the strategy that best accounted for their responses and the strategy that worst captured their responses. Experiment 1.

Figure 3.3 shows the overall picture, with an analysis in which data are collapsed across tests: All averaging strategies better captured participants' responses than Bayesian and matching accounts did. In particular, the average between the prior probability and the likelihood had a ratio of 15, while the average between the likelihood and the posterior probability had a ratio of 5.09. All other strategies had a ratio < 1, meaning that they were more predictive of responses which participants did not provide than of estimates which participants actually gave.

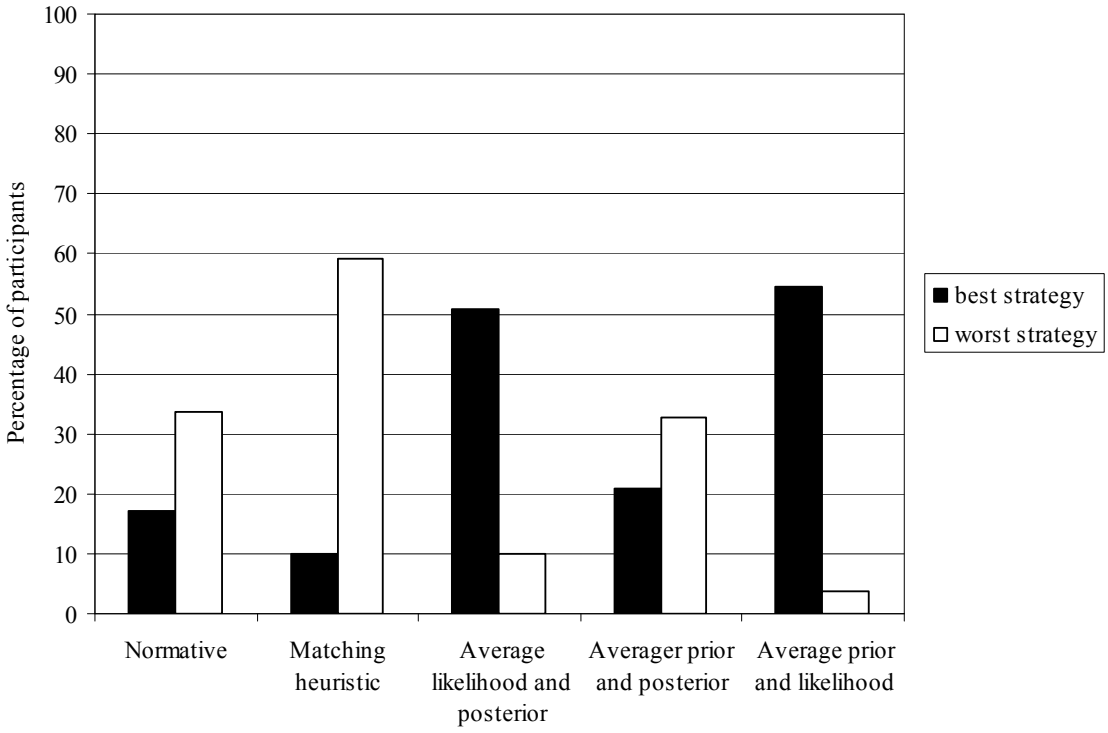


Figure 3.3. Distributions of participants, collapsing the responses of the four tests, as a function of the strategy that best accounted for their responses and the strategy that worst captured their responses. Experiment 1.

Experiment 2

Experiment 2 was set up to address the issue of a possible effect of the format of the presented information on participants' confidence after a "yes" and after a "no". In Bayesian terms, the ability to differentiate (or equate, in case of symmetric tests) the "yes" and the "no" answers in terms of the different (equal) informativeness they hold entails the computation of both the LR relative to the "yes" answer and the LR of the "no" answer. We hypothesized that people might encounter more difficulties in considering the latter when evaluating "yes" and "no" answers to the same question. In our task, the computation of the LR relative to the "no" answer implies the ability to figure out the probabilities of non-occurrence of a feature in the two groups. This might be harder than taking into account the probabilities of occurrence because of the well-known difficulty to process negative information than positive information (e.g., Hearst, 1991; Van Wallendael, 1995; Wason, 1959, 1961; we shall further deepen this issue in the next chapter), and also because the probabilities relative to the presence of the features are given, while those relative to the absence are not given.

Accordingly, it might be that when adding to the probabilities of the presence of the features the probabilities of their absence participants would be more sensitive to the actual informativeness of "yes" and "no" answers. We tested this hypothesis in Experiment 2. Participants in Experiment 1 were presented with *p*s (E | H), where "H" stands for both the hypotheses (i.e., both Gloms and Fizos), while in Experiment 2 they received both *p*s (E | H) and *p*s (¬E | H) (see Appendix H, the procedure was drawn from Experiment 2 by Cherubini, Rusconi, Russo, & Crippa, submitted, see Chapter 4).

Participants

Ninety-four undergraduate students at the University of California, San Diego (64 female, 30 male, mean age = 20.2 years, range: 17-28 years; 65 were native speakers of English) took part in the study in exchange for course credit.

Materials and procedure

Design, materials, instructions and procedure were exactly the same as in Experiment 1, with the exception of the addition of the probabilities of the absence of the features beside the probabilities of their presence (see Appendix G). For instance, when presenting to participants test 4 we gave them both the 50-2 percentage combination, indicating the probabilities of the presence of the feature in Gloms and Fizos, and its complement, the 50-98 combination, indicating the probabilities of the absence of the same feature in the two groups.

Results and discussion

A/symmetry of confidence after a "yes" and after a "no". We used the same recoding of participants' estimates used in Experiment 1 (see also McKenzie 2006; Slowiaczek et al., 1992). A series of paired *t*-tests showed that, contrary to Experiment 1, participants' mean estimates after a "yes" did not significantly differ from mean estimates after a "no". Indeed, for test 1 (65-35 feature), the mean estimate after the "yes" (M = 61.8, SD = 14.53) was not significantly different from the mean estimate after the "no" (M = 58.04, SD = 16.01), $t(93) = 1.36$, $p = .18$ (95% CI [-1.74, 9.26]) (Figure 3.4, top left panel). In a similar vein, for test 2 (85-15 percentage combination), participants' mean estimate after the "yes" (M = 75.34, SD = 19.35) did not differ from the mean estimate after the "no" (M

= 71.99, *SD* = 22.96), *t*(93) = 1.05, *p* = .30 (95% CI [-3.00, 9.69]) (Figure 3.4, top right panel).

Participants exhibited symmetry of confidence when receiving the answers to the asymmetrically disconfirming question (98-50 feature). A one-sample *t*-test revealed that the mean difference of -.63 (*SD* = 33.47) between confidence after the "yes" and confidence after the "no" was significantly less than the normative 30, *t*(93) = -8.87, *p* = .001, *d* = .88 (95% CI [-37.49, -23.78]). Indeed, as shown by a paired *t*-test, the mean estimate after the "yes" answer (*M* = 65.59, *SD* = 20.31) was not significantly different from the mean estimate after the "no" answer (*M* = 66.23, *SD* = 30.07), as shown by a paired *t*-test, *t*(93) = - .18, *p* = .85 (95% CI [-7.49, 6.22]) (Figure 3.4, bottom left panel). Finally, participants perceived a difference between "yes" and "no" to the asymmetrically confirming query (50-2 feature), but a one-sample *t*-test revealed that the mean perceived difference of 13.38 (*SD* = 36.3) was significantly less than the normative difference of 30, *t*(93) = -4.44, *p* = .001, *d* = .46 (95% CI [-24.06, -9.19]). A paired *t*-test showed that the mean estimate after the "yes" (*M* = 72.63, *SD* = 28.05) was significantly different from the mean estimate after the "no" (*M* = 59.25, *SD* = 22.72), *t*(93) = 3.57, *p* = .001, *d* = .53 (95% CI [5.94, 20.81]) (Figure 3.4, bottom right panel).

Hence, participants benefited from receiving fully explicit distribution of probabilities (i.e., both the probabilities of the presence and the probabilities of the absence of the features inquired about) when estimating the confidence in the hypothesis after a "yes" and after a "no" to symmetric queries (i.e., tests 1 and 2). Indeed, compared to Experiment 1, they exhibited more symmetry of confidence. The performance was more similar to the normatively expected behavior also when evaluating the diagnosticity of answers to asymmetric questions. The answers to the asymmetrically disconfirming query (about the 98-50 feature) were perceived as almost equally diagnostic, but, differently from Experiment 1, there was a tendency to perceive a greater weight of the "no" vs. the "yes", in line with the normative direction of the asymmetry (see Figure 3.4, bottom left panel). When judging the informativeness of answers to the asymmetrically confirming question (50-2 percentage combination) there was a tendency to perceive a greater difference between the two answers' informativeness than in Experiment 1 (mean difference of 13.38 in Experiment 2 vs. mean difference of 10.4 in Experiment 1).

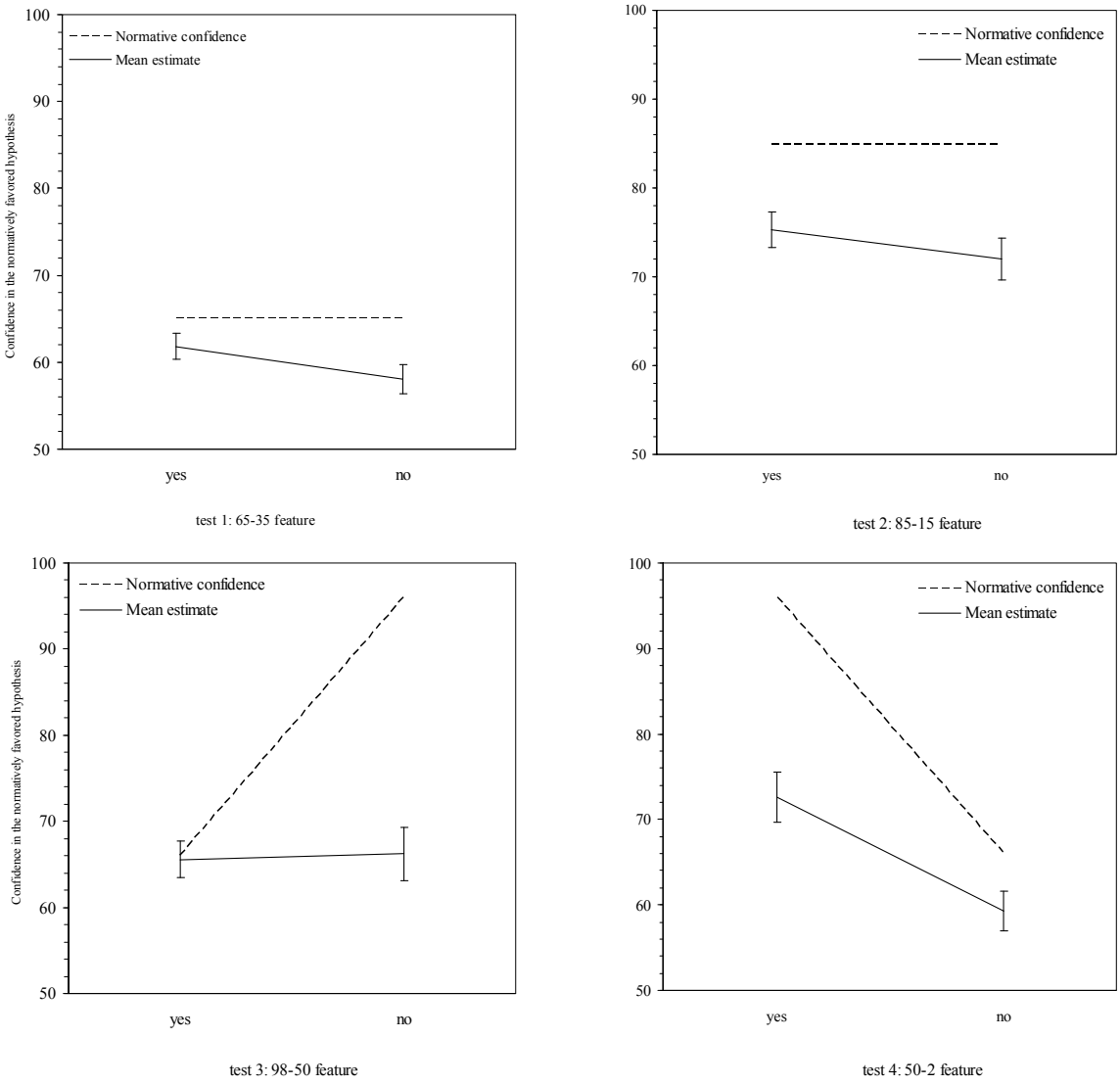


Figure 3.4. Confidence in the hypothesis favored by the evidence. The participants' mean estimates are compared with the normatively expected confidence. For participants' mean estimates, standard error of the mean (SEM) bars are also shown. Experiment 2.

Comparing predictors of insensitivity/oversensitivity to answer diagnosticity. As in Experiment 1 we compared the predictive power of different strategies which might underlie the judgment of differently informative answers. The accounts as well as their predictions are the same as those illustrated in Table 3.2 for Experiment 1, except for the averaging between the likelihood and the posterior probability, whose predictions changed for the "no" answers, as shown in Table 3.5.

Strategies	Answer	Predictions	test #1: 65%-35%	test #2: 85%-15%	test #3: 98%-50%	test #4: 50%-2%
Matching heuristic	*yes*	confusion of p (H | D) with p (D | H)	65	85	98	50
	no	confusion of p(H | ¬D) with p (D | H)	35	15	50	2
Average between the likelihood and the posterior probability	*yes*	p (D | H) ≤ estimate ≤ p (H | D)	[65, 65]	[85, 85]	[98, 66]	[50, 96]
	no	p (¬D | H) ≤ estimate ≤ p (H | ¬D)	[65, 65]	[85, 85]	[50, 96]	[98, 66]
Average between the prior and the posterior probabilities	*yes*	p (H) ≤ estimate ≤ p (H | D)	[50, 65]	[50, 85]	[50, 66]	[50, 96]
	no	p (H) ≤ estimate ≤ p (H | ¬D)	[50, 65]	[50, 85]	[50, 96]	[50, 66]
Average between the prior probability and the likelihood	*yes*	p (H) ≤ estimate ≤ p (D | H)	[50, 65]	[50, 85]	[50, 98]	[50, 50]
	no	p (H) ≤ estimate ≤ p (¬D | H)	[50, 65]	[50, 85]	[50, 50]	[50, 98]

Table 3.5. The strategies (and their predictions) pitted against one another in Experiment 2. For each test that we used point values or intervals resulting from each account's predictions are provided. "H" stands for the normatively favored hypothesis (i.e., "Glom" when the answer was "yes", "Fizo" when the answer was "no").

In Experiment 1, with respect to this strategy, we hypothesized that people could average the likelihoods they received, which were always *p*s (E | H), with "H" meaning either "Gloms" or "Fizos", and the normative posterior probabilities. For this reason, we expected that the estimate of a participant using this strategy should fall between p (E | H) and p (H | E) also when the answer was "no", that is when the feature did not occur (i.e., ¬E). Differently, in Experiment 2, we hypothesized that supplying participants with fully explicit probabilistic information not only about the presence, but also about the absence of the features should lead them to consider the probability of non-occurrence of the feature under the working hypothesis (i.e., p (¬E | H)) when they receive the "no" answer. Accordingly, when the feature under consideration was absent, the prediction was of an average between p (¬E | H) and p (H | ¬E) (see Table 3.5).

Of course, we did not apply the same reasoning as to the matching heuristic, because its predictions when the feature is absent (i.e., ¬E) are diagnostic to distinguish it from the inverse fallacy (Villejoubert & Mandel, 2002).

We performed the same analyses described as for Experiment 1. Accordingly, once determined the predicted intervals of confidence for each account (see Table 3.6, the only difference compared to Table 3.4 is relative to the predictions of the strategy of averaging the likelihood and the posterior probability), we examined the ratio of the percentage of participants for whom a specific strategy was the best account for their responses to the percentage of participants for whom the same strategy was the worst predictor.

Strategies	Answer	Predictions	test #1: 65%-35%	test #2: 85%-15%	test #3: 98%-50%	test #4: 50%-2%
Normative	*yes*		$60 < x < 70$	$80 < x < 90$	$61 < x < 71$	$91 < x \leq 100$
	no		$50 < x < 80$	$80 < x < 90$	$91 < x \leq 100$	$61 < x < 71$
Matching heuristic	*yes*	confusion of p (H | D) with p (D | H)	$60 < x < 70$	$80 < x < 90$	$93 < x \leq 100$	$45 < x < 55$
	no	confusion of p(H | ¬D) with p (D | H)	$20 < x < 50$	$10 < x < 20$	$45 < x < 55$	$0 \leq x < 7$
Average between the likelihood and the posterior probability	*yes*	p (D | H) ≤ estimate ≤ p (H | D)	$60 < x < 70$	$80 < x < 90$	$66 \leq x \leq 98$	$50 \leq x \leq 96$
	no	p (¬D | H) ≤ estimate ≤ p (H | ¬D)	$60 < x < 70$	$80 < x < 90$	$50 \leq x \leq 96$	$66 \leq x \leq 98$
Average between the prior and the posterior probabilities	*yes*	p (H) ≤ estimate ≤ p (H | D)	$50 \leq x \leq 65$	$50 \leq x \leq 85$	$50 \leq x \leq 66$	$50 \leq x \leq 96$
	no	p (H) ≤ estimate ≤ p (H | ¬D)	$50 \leq x \leq 65$	$50 \leq x \leq 85$	$50 \leq x \leq 96$	$50 \leq x \leq 66$
Average between the prior probability and the likelihood	*yes*	p (H) ≤ estimate ≤ p (D | H)	$50 \leq x \leq 65$	$50 \leq x \leq 85$	$50 \leq x \leq 98$	$45 < x < 55$
	no	p (H) ≤ estimate ≤ p (¬D | H)	$50 \leq x \leq 65$	$50 \leq x \leq 85$	$45 < x < 55$	$50 \leq x \leq 98$

Table 3.6. The intervals of confidence pitted against one another and with participants' estimates in Experiment 2.

It turned out that, differently from Experiment 1, in the symmetric tests the average between the likelihood and the posterior probability performed worse in capturing participants' responses than the other two averaging accounts, which, for the symmetric queries, lead to the same predictions. In particular, for test 1 (65-35 feature), both the average between the prior and the posterior probabilities and the average between the prior probability and the likelihood had a best/worst ratio of 1.94. The Bayesian strategy had a ratio of 1.27, while the average between the likelihood and the posterior had a ratio of 1.04. The worst account was the matching heuristic (.27) (see Figure 3.5, top left panel). A similar pattern emerged when considering test 2 (85-15 feature). The average between the prior and the posterior probabilities and the average the prior probability and the likelihood were the best predictors (ratio of 2.62), followed by the Bayesian strategy and the average between the likelihood and the posterior (both 1.12). Again the matching strategy was the worst predictor for more participants than those for whom it was the best account (ratio of .04) (Figure 3.5, top right panel).

As in Experiment 1, for both the asymmetric tests, the only strategies which had a ratio > 1 were the average between the likelihood and the posterior probability and the average between the prior probability and the likelihood. Specifically, for test 3 (98-50 percentage combination), the best account turned out to be the average between the prior probability and the likelihood (ratio: 5.5), followed by the average between the likelihood and the posterior probability (3.07). The other strategies had more participants for whom they were the worst account than participants for whom they represented the best account (average between prior and posterior probabilities: .74; Bayesian: .73; matching heuristic: .47) (Figure 3.5, bottom left panel). Similarly, for test 4 (50-2 feature), the average between the prior probability and the likelihood was the best predictor (ratio: 2.62), followed by the average between the likelihood and the posterior probability (1.26), while the other accounts were more worst than best predictors (Bayesian: .98; average between the prior and the posterior probabilities: .76; matching heuristic: .14) (Figure 3.5, bottom right panel).

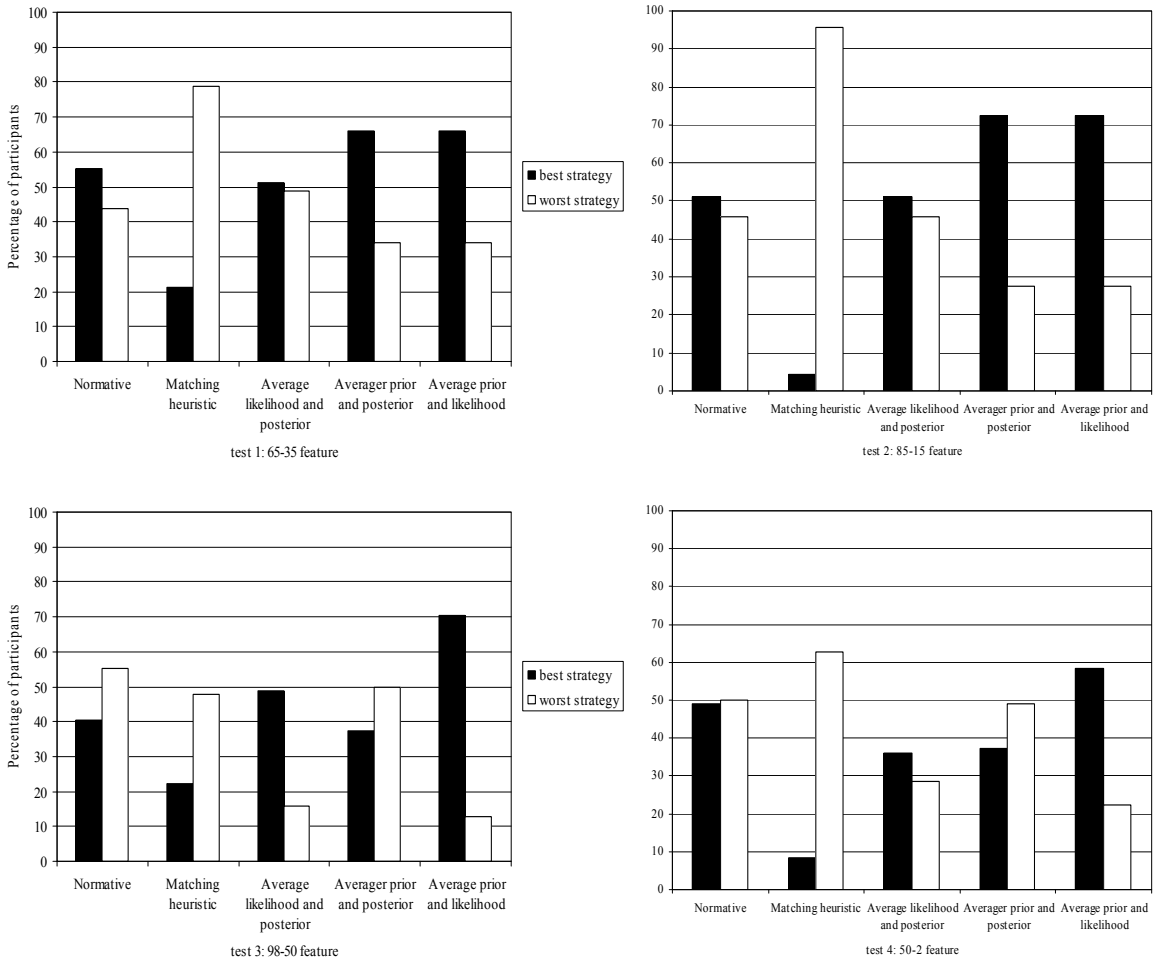


Figure 3.5. Distributions of participants, for each test, as a function of the strategy that best accounted for their responses and the strategy that worst captured their responses. Experiment 2.

Overall, as shown in Figure 3.6, it turned out that, collapsing across tests, averaging the prior probability and the likelihood was the strategy that for 74.47% of participants best accounted for their responses, while for none of the participants it was the worst account. The average between the likelihood and the posterior probability had a ratio of 4.29, while the Bayesian strategy had a ratio of 1.45. Finally, the matching heuristic was the worst predictor (ratio: .23): for 69.15% of participants it represented the strategy that worst accounted for their responses.

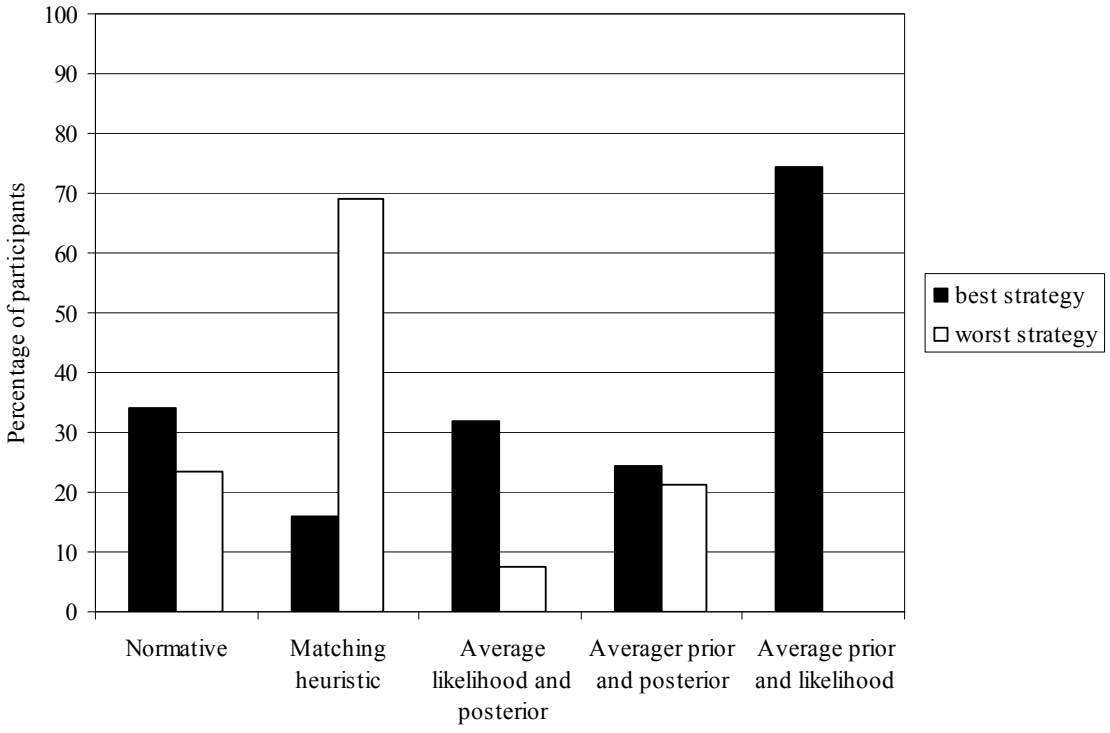


Figure 3.6. Distributions of participants, collapsing the responses of the four tests, as a function of the strategy that best accounted for their responses and the strategy that worst captured their responses. Experiment 2.

General discussion

Optimal selection of information does not guarantee optimal revision of prior beliefs. Indeed, asking the most useful questions does not necessarily prevent people from misweighing the evidence acquired by means of those queries. Vice versa an unbiased information processing does not imply that people have exposed themselves in a fairly fashion to the information available, accordingly people might inappropriately confirm or revise their pre-existing expectancies also when a correct evaluation of the collected data has been accomplished.

The two experiments we have presented in this chapter focused on the evaluation stage of hypothesis development (e.g., Klayman, 1995, McKenzie, 2004). We attempted to further address the issue of the relative insensitivity to answer diagnosticity found in previous research with abstract material (McKenzie, 2006; Skov & Shrman, 1986; Slowiaczek et al., 1992), by pitting against one other four strategies which might be used when evaluating the informativeness of answers in abstract tasks. In doing so, we took

into account the structure of the questions from which the answers came (i.e., their symmetry vs. asymmetry, which is tantamount to considering the equal or different evidential strength of the answers), and the presentation format of the probabilities of the features inquired about (i.e., the way the likelihoods are given to participants).

Our findings add to the existing literature with respect to three points, which will be discussed in the following pages. First, they show that insensitivity to differential answer diagnosticity should be regarded not only as "insensitivity" but also as "oversensitivity", that is, people fail not only to appreciate that two answers to the same question convey a different amount of information, but also that two answers can weigh equally. In other words, the type of question asked can influence the way its answers are evaluated. Second, and related to the first point, they suggest that taking into account the type of question asked can give insights on how hypothesis testing and hypothesis evaluation combine and lead to confirmation bias (or to its mitigation). Third, our data do not provide supporting evidence to the hypothesized tendency to assess answer diagnosticity as though it was question diagnosticity (Slowiaczek et al., 1992). Indeed, we found evidence that averaging strategies can better account for participants' responses. Specifically, people would over-rely on the information they receive, that is the prior probability and the likelihood under the working hypothesis.

With respect to the issue of how people use information compared to the Bayesian criterion, we found evidence of answer oversensitivity in Experiment 1, in which participants tended to perceive as differentially diagnostic answers of equal strength, as those to tests 1 (65-35 feature) and 2 (85-15 feature). This result suggests that the relative insensitivity to answer diagnosticity can be better conceived as a more general failure in information use, that can occur in terms of either insensitivity or oversensitivity depending on the question asked. Furthermore, this finding runs counter to the interpretation of insensitivity to answer diagnosticity provided by Slowiaczek et al. (1992), who argued that people would use the "feature-difference heuristic" (Nelson, 2005, footnote 2; Nelson et al., 2010; Slowiaczek et al., 1992; see also Chapter 1) which is useful for assessing question informativeness. Indeed, the difference in likelihoods was the same for "yes" and "no" answers (.3 in test 1 and .7 in test 2), thus participants should have judged as equally diagnostic the two answers.

As to the asymmetric questions (i.e., test 3 about the 98-50 feature and test 4 about the 50-2 feature), the results showed the pattern of relative insensitivity that previous

research has already pointed out. However, it should be noted that, in both experiments, participants were less sensitive to the differential diagnosticity of the answers to the asymmetrically disconfirming question (98-50 feature) compared to the answers to the asymmetrically confirming question (50-2 feature)[18]. Specifically, in Experiment 1 there was a tendency to weight more the "yes" answer than the "no" answer to the asymmetrically disconfirming query (98-50 feature), which is opposite to the normative direction. By contrast, participants perceived a difference, in the normatively expected direction, between "yes" and "no" answers to the asymmetrically confirming query (50-2 feature), even though insufficiently. In Experiment 2, participants perceived the two answers as almost equally diagnostic when the question was asymmetrically disconfirming, while they exhibited an insufficient asymmetry of confidence when the question was asymmetrically confirming.

This differential magnitude of insensitivity to answer diagnosticity when the question is asymmetrically confirming vs. asymmetrically disconfirming cannot be explained by a confusion of the assessment of answer diagnosticity with the assessment of question usefulness (Slowiaczek et al., 1992), that predicts almost the same estimates of confidence after all the answers to the asymmetric tests. Indeed, the difference in likelihoods is always .48 for "yes" and "no" answers of both tests.

Taking into account the type of question asked is important to relate insensitivity/oversensitivity to answer diagnosticity to confirmation bias, which is a combination of a testing strategy and an evaluation strategy (e.g., McKenzie 2004, 2006; Klayman, 1995; Poletiek, 2001). Slowiaczek et al. (1992) argued that "symmetrical questions (70-30, 20-80) are not prone to the inferential errors we document, because "yes" and "no" answers are equally diagnostic" (Slowiaczek et al., 1992, p. 402). In light of our results, we can conclude that inferential errors might occur also when symmetric questions are asked. We found that people weighed more the "yes" answer than the "no" answer to a symmetric query (Experiment 1). This "feature-positive effect" (e.g., Hearst & Wolff, 1989; Klayman, 1995; McKenzie, 2006; Newman, Wolff, & Hearst, 1980; see

[18] A greater insensitivity to answer diagnosticity when the "yes" and the "no" come from an asymmetrically disconfirming question rather than from an asymmetrically confirming question has been found also by Slowiaczek et al. (1992, Experiment 1A). Indeed, a close scrutiny of the data from their Experiment 1A (see Slowiaczek et al., 1992, Table 2, p. 396) reveals that when the percentage combinations were either 50%-10%, or 10%-50%, that is when the tests were asymmetrically confirming, the estimated difference between the informativeness of "yes" and "no" was 11%, while when the answers came from the asymmetrically disconfirming tests, the estimated difference decreased to 4% for the 50%-90% combination and to 2% for the 90%-50% combination (the normative difference was of 19% for all percentage combinations).

also Cherubini et al., submitted, which we shall present in the next chapter) combined with a preference for symmetric questions which are positive (positivity/negativity and symmetry/asymmetry are independent characteristics of the questions, see Cherubini et al., 2010) would lead to confirmation bias, as described by Klayman (1995) (see also McKenzie, 2006).

Insensitivity to answer diagnosticity might lead to different consequences in terms of confirmation bias depending also on the type of asymmetric question asked. As Slowiaczek and co-workers noted (see also McKenzie, 2006; Klayman, 1995), a combination of the preference for "extreme" questions (i.e., asymmetrically disconfirming questions, see Cherubini et al., 2010) and a failure to perceive that the "no" answer to this kind of questions is more diagnostic than the "yes" answer (in other words, a combination of extremity and insensitivity) leads to confirmation bias. However, confirmation bias might be conceived not only in terms of committing Type I errors (i.e., false positives errors), which are more likely when asking asymmetrically disconfirming questions (Cherubini et al., 2010; Trope & Liberman, 1996), but also in terms of incurring in Type II errors (i.e., false negative errors), which are more likely when asking asymmetrically confirming questions (Cherubini et al., 2010; Trope & Liberman, 1996). The latter, that is confirmation bias in terms of preferring asymmetrically confirming questions, has been found by Trope & Thompson (1997), Cameron & Trope (2004) and Poletiek & Berndsen (2000) in more contextualized tasks than those used by us and by Skov & Sherman (1986) and Slowiaczek et al. (1992). A preference for this kind of questions combined with the relative insensitivity to the differential diagnosticity of the answers that we found would make less likely a confirmation bias (in this case, defined as the maximization of the *strength* of the evidence confirming the focal hypothesis), because the "yes" answer would be judged less diagnostic than actually it is.

Hence, while insensitivity to answer diagnosticity leads to confirmation bias (in this case, defined as the maximization of the *probability* of occurrence of the confirming evidence) when combined with an asymmetrically disconfirming testing strategy, it turns out that it weakens the confirmation bias (conceived as the maximization of the *strength* of the confirming evidence) when combined with an asymmetrically confirming testing strategy. However, this "debiasing effect" of insensitivity to answer diagnosticity when the questions are asymmetrically confirming, as well as the reciprocal confirmation-bias effect when the questions are asymmetrically disconfirming, might hold only when

abstract materials are used, because familiar materials increase sensitivity to answer diagnosticity (McKenzie, 2006).

If the confusion of the assessment of answer diagnosticity with the assessment of question usefulness advanced by Slowiaczek et al. (1992) does not seem to explain the insufficient sensitivity to answer diagnosticity in abstract tasks, by which strategy can account for it? To address this issue, we compared four strategies. Specifically, we pitted against one another the matching heuristic (already tested in the inductive domain by Villejoubert & Mandel, 2002) and three averaging accounts: The average between the likelihood and the posterior probability, the average between the prior and the posterior probabilities, and the average between the prior probability and the likelihood. We found further evidence against a primary role of matching heuristic in inductive reasoning (Villejoubert & Mandel, 2002). Indeed, matching was the only strategy that, both in Experiment 1 and in Experiment 2 and for any combinations test-answer, had a ratio of the percentage of participants for whom it was the best account for their responses to the percentage of participants for whom it was the worst account inferior to 1. This suggests that participants are not determined only by the superficial properties of the task, such as the distributions of probabilities they received.

By contrast, as shown in Figures 3.3 and 3.6, the best predictor of participants' responses was the average between the prior probability and the likelihood, which had always (in both experiments and for any tests) a best/worst ratio > 1. This finding suggests that, contrary to the interpretations in terms of a failure to consider base rates, which Skov & Sherman (1986) hinted at, and more generally in terms of the use of representativeness (Slowiaczek et al., 1992), participants seemed well aware of the prior probabilities when estimating answer diagnosticity. Not only the prior probabilities did affect their estimates, but also the only other information they were given, that is the likelihood. It is not surprising, indeed, that the average between the prior probability and the likelihood turned out to be a more powerful predictor in Experiment 2 than in Experiment 1 (in Experiment 1 it was the best predictor for 54.55% of participants and the worst for 3.64 of them, whereas in Experiment 2 it represented the best account for 74.47% of participants, while being the worst account for none of the participants). Indeed, in Experiment 2, but not in Experiment 1, participants were presented with fully explicit information about the likelihoods of occurrence *and* the likelihoods of non-

occurrence of the features to inquire about, thus they were more likely to be influenced by the information they were given than in Experiment 1.

This finding is in keeping with the already found descriptive power of the strategy of averaging priors and likelihoods (McKenzie, 1994). Furthermore, it shows that people correctly identify which pieces of information are relevant to the task, but they fail in integrating them in a Bayesian fashion.

Finally, we found that the presentation format affected participants' responses, even though not enough to overcome completely the relative insensitivity/oversensitivity to answer diagnosticity. Indeed, the presentation of the distributions of probabilities relative to both the presence and the absence of the features inquired about improved participants' judgments in all tests in Experiment 2. For both symmetric questions, the difference in estimated diagnosticity between "yes" and "no" ceased to be significant, even though it was not null. For the asymmetrically disconfirming question (98-50 feature), there was a slight tendency to perceive the greater informativeness of the "no" answer compared to the "yes" answer. Finally, for the asymmetrically confirming question (50-2 feature), the insensitivity was less pronounced with respect to Experiment 1. This suggests that people's difficulty to normatively revise their initial beliefs in light of different answers to the same question might reside not only in overreliance on given information, but also in figuring out the likelihoods relative to the non-occurrence of the features inquired about.

We have addressed the issue of how people interpret new information given within an abstract task in a particular subset of situations, namely when the question/test allows only dichotomous answers, as "yes" and "no" are. However, how about the given pieces of evidence are more complex? Is people's sensitivity to the informativeness of the incoming data somehow affected by other, formally irrelevant, properties of the data? In the next chapter, we shall work out in depth the issue emerged also in the experiments we have presented in this chapter of whether and to what extent people are driven by the presence of the incoming information more so than by its absence.

Chapter 4

A feature-positive effect in hypothesis evaluation[19]

Feature-positive effects refer to the predisposition of human beings and other animals to pay more heed to the occurrences of stimuli rather than to their non-occurrences (e.g., Jenkins & Sainsbury, 1969; Newman et al., 1980). At the sensorial level, presence is detected faster than absence. Features that are present in the environment are transduced into electrophysiological signals faster than those signaling the absence of important features (e.g., hunger or thirst). In some domains, this peculiarity persists at the cognitive level: It is a quite common experience that in familiar environments, we often realize that a new object is present faster than we realize that an old object is absent. It was conjectured that feature-positive effects are an adaptation to a typical information pattern, whereby the occurrences of particular features are relatively rare compared to their non-occurrences, and thus they are, from a very general perspective, more informative (McKenzie & Mikkelsen, 2007; Newman et al., 1980). Once consolidated, this tendency to overrate the presence of stimuli may also generalize to those contexts in which the presence of certain stimuli does not necessarily convey more information than their absence. The present experiments investigate whether and to what extent people overrate the informational value of present vs. absent features when they evaluate alternative hypotheses—that is, when they determine which of two mutually exclusive hypotheses is most likely in light of available data. This issue is important for cognitive psychology, because many scholars argue that *positive testing*, a quite common and spontaneous hypothesis-testing strategy, might result in confirmation biases if combined with feature-positive effects (e.g. Klayman, 1995; McKenzie, 2004, 2006). Yet, to the best of our knowledge, no direct empirical evidence has ever corroborated the idea that present clues are rated as more important than absent clues when alternative hypotheses are evaluated.

[19] This chapter is made up of materials which appear in Cherubini, Rusconi, Russo, & Crippa (submitted). Missing the dog that failed to bark in the nighttime: On the overestimation of occurrences over non-occurrences in hypothesis testing. The authors wish to thank Katya Tentori for her insightful suggestions.

Overview of previous literature

Feature-positive effects were described in several domains. The studies of discrimination learning, involving many cross-species experiments (e.g., from animals, such as pigeons, to young children and adults) and employing a range of stimuli, procedures and experimental settings (e.g., Hearst & Wolff, 1989; Newman et al., 1980), have shown that the ability to discriminate between two stimuli that differ only by the presence or absence of a feature is acquired more rapidly and correctly when the feature is present on positive trials rather than on negative trials. These results indicate that the presence of a feature directs attention more so than its absence. Similar effects were observed at increasingly higher levels of cognitive processing. The presence of characteristics is more relevant than their absence in the learning of concepts. The acquisition of a concept is easier for people when they receive positive instances (i.e., information about what the concept is) rather than negative instances (i.e., information of what it is not) (Bourne & Guy, 1968; Hovland & Weiss, 1953; Klayman, 1995; Nahinsky & Slaymaker, 1970). In probability learning, people tend to make their predictions on the basis of the relative frequency of the occurrence of different categories of stimuli, instead of on the basis of the actual probability of each type of stimulus because the latter would require the accurate recall of trials in which the stimulus did not occur (Estes, 1976). In yet another domain, when evaluating two-way contingency tables, people weigh the co-occurrences of stimuli more than the instances in which one or both of the stimuli is absent, a phenomenon labeled *cell weight inequality* (e.g., Beyth-Marom, 1982; Jenkins & Ward, 1965; Kao & Wasserman, 1993; Mandel & Lehman, 1998). Although it might be argued that this tendency is normatively adequate when the stimuli are rare (McKenzie & Mikkelsen, 2007), in other contexts it inflates illusory correlations (e.g., Mandel & Lehman, 1998; Smedslund, 1963).

In hypothesis development, which is the focus of interest of the present contribution, it is well known that, as we have already described in previous chapters, when gathering information for checking whether a hypothesis is true or false, there is a moderate to strong tendency to adopt a positive testing strategy (Baron et al., 1988; Cherubini et al., 2010; Klayman, 1995; Klayman & Ha, 1987; Skov & Sherman, 1986; Slowiaczek et al., 1992; Snyder & Swann, 1978; Wason, 1960). Positive testing, in its current understanding, affects the *gathering*, as opposed to the *evaluation*, of information. We remind to the reader that it consists of a tendency to preferentially look for the occurrence

of features that are more probable when the tested hypothesis is true than when it is false. The occurrence of those features strengthens (namely, inductively confirms) the tested hypothesis, whereas their non-occurrence weakens (i.e., inductively falsifies) it. It is easy to see the possible consequences of a feature-positive effect in the evaluation stage of hypothesis testing, for individuals adopting positivity as a strategy in information gathering: First, features whose occurrence might verify the hypothesis are searched for; second, if such confirming features indeed occur, they are attended and considered; conversely, if they do not occur, the corresponding falsification of the hypothesis could be neglected or underestimated. The result could be the systematic, improper apportionment of excessive confidence in the truth of the tested hypothesis, namely a *confirmation bias* (Klayman, 1995; McKenzie, 2004, 2007; Nickerson, 1998).

The empirical evidence for or against the occurrence of a feature-positive effect in the evaluation stage of hypothesis testing is scant. Fischhoff and Beyth-Marom listed the effect as a typical deviation from a correct Bayesian evaluation of a hypothesis: "In principle, people can ignore the likelihood ratio just as well as the base rate [...]. This may happen, for example, when the datum [...] reports a non-occurrence. A classic example of the latter is Sherlock Holmes's observation (Doyle, 1974) that his colleague, Inspector Gregory, had not considered the significance of a dog failing to bark when an intruder approached." (Fischhoff & Beyth-Marom, 1983, p. 246). However, the authors did not report empirical evidence in support of the existence and magnitude of such a tendency apart from the anecdotic reference to Arthur Conan Doyle's tale. Screening the relevant literature, we found many references to the possibility that non-occurrences are underestimated in the evaluation stage of hypothesis testing (e.g., in Klayman, 1995; McKenzie, 2004, Nickerson, 1998), but the empirical evidence is very scant. In their work on belief revisions, which was mainly focused on how people use answers to questions concerning the presence of features in individuals from a given population, Slowiaczek et al. (1992) provided some evidence of a feature-positive effect, but not consistently across studies. In Experiment 1A, they found a tendency to weigh "yes" answers (indicating that a feature is present) more than "no" answers (indicating that a feature is absent), regardless of the actual informativeness of the answers. Yet, in Experiments 2A, B, and C, this effect was present only when participants estimated the composition of a population from a sample. Also in the two experiments we have

presented in the previous chapter, we found supporting for a tendency to give more weight to the "yes" answer than to the "no" answer.

The only other empirical investigation that we managed to find that directly and specifically tested the feature-positive effect in hypothesis evaluation is Christensen-Szalansky and Bushyhead's 1981 study on medical diagnosis in a real clinical setting: "This study also examined the physicians' ability to estimate the predictive value of an "absent symptom", since the absence of a symptom also can be helpful in assigning a diagnosis. Past psychological research has suggested that people do not efficiently process the "absence of cues" (Bourne & Guy, 1968; Hovland & Weiss, 1953; Nahinsky & Slaymaker, 1970)." (Christensen-Szalansky & Bushyhead, 1981, p. 931; the studies that the authors mention in this sentence concern feature-positive effects in rule and concept learning, but not in hypothesis evaluation). Actually, Christensen-Szalansky and Bushyhead failed to find a significant underestimation of the diagnostic strength of absent symptoms, but they were very cautious about their negative finding: "the realism of the study reduced the experimenters' control of the presence of correlated symptoms. For example, if the absence of symptom X always occurred with the presence of important symptom Y, perhaps physicians' apparent "use" of the absent symptom was simply an artefact due to this correlation. A more controlled experiment is needed to support these results" (p. 934). We did not find any more controlled experiments on this topic in later research.

Basic formal concepts about hypothesis testing

From a logical perspective, inductive hypothesis testing and belief update are mostly viewed (but see Cohen, 1977) as a change in the epistemic probability *p* that a hypothesis "H" is true (as opposed to false, corresponding to the probability that its complement, "¬H", is true) after acquiring a piece of evidence "E", with respect to the probability that "H" was true before "E" was acquired. A widespread formal method of belief update is Bayes' rule (see Chapter 1). Given the prior probability of "H" and its posterior probability following the receipt of "E" (computed by Bayes' rule), it is possible to formally estimate the informational value of "E" in terms of Shannon's (1948) entropy—that is, in bits (see Chapter 1). The information gain associated with a body of evidence "E", namely ΔI_E is the difference between initial entropy and entropy after "E" has been taken into account:

$$\Delta I_E = \{[p(H) \times \log_2 1 / p(H)] + [p(\neg H) \times \log_2 1 / p(\neg H)]\} - \{[p(D) \times ((p(H \mid D) \times \log_2 1 / p(H \mid D)) +$$
$$+ (p(\neg H \mid D) \times \log_2 1 / p(\neg H \mid D)))] + [p(\neg D) \times ((p(H \mid \neg D) \times \log_2 1 / p(H \mid \neg D)) +$$
$$+ (p(\neg H \mid \neg D) \times \log_2 1 / p(\neg H \mid \neg D)))]\}$$

For example, in a situation in which we are torn between two complementary and well-defined hypotheses, with p (H) = p (¬H) = .5, initial entropy is $\log_2(2) = 1$ bit. Two events occur, E_1 and E_2. The former has p (E_1 | H) = .8 and p (E_1 | ¬H) = .4, with LR = 2, thus resulting, by Bayes' rule above, in a posterior odds ratio of 2:1 in favor of H. The latter has p (E_2 | H) = .3 and p (E_2 | ¬H) = .1, with LR = 3, that, when applied to the previously revised odds, results in a final posterior odds ratio of 6:1 in favor of H, equivalent to p (H | E) = 6/7 = .86 (where E stands for the whole body of evidence, including both E_1 and E_2; the order of their receipt does not affect the computation). The final entropy of the system denoted by the two hypotheses is $.86\log_2(.86) + .14\log_2(.14) = .59$ bits. The whole body of evidence E has thus conveyed $\Delta I_E = 1 - .59 = .41$ bits. The ΔI_E is a convenient quantitative measure for estimating the amount of information conveyed by a set of clues.

From a formal standpoint, it does not matter whether information is conveyed by the presence of an attribute in a situation or by its absence. A highly likely occurrence shifts the belief towards a hypothesis exactly as the non-occurrence of a highly unlikely event, and vice versa. Accordingly, testing whether the occurrence of features affects belief revisions more so than their absence is equivalent to testing whether people, with regard to their spontaneous belief revisions, are biased by a formally irrelevant aspect of the situation. Such bias might have relevant practical consequences in professions in which accurate belief revision is critically important, for example, for judges who have to infer a verdict from different clues (e.g., Wells & Lindsay, 1980) or physicians who must formulate a diagnosis. For instance, in a patient with symptoms of hyperthyroidism, the assessment of normal ocular objectivity conveys the same diagnostic value as the reading of the absence of exophthalmos: Both clues should lead a physician towards a diagnosis of a form of non-Basedow thyreopathy (Scandellari, 2005). If physicians systematically underestimate the relevance of absent signs, however, the diagnostic importance of the

absence of exophthalmos would be underestimated, resulting in weaker than warranted diagnostic hypotheses.

Overview of the three experiments

Goal

In three paper-and-pencil experiments sharing the same experimental design and similar procedures, we investigated whether and to what extent people overestimate the importance of present features in contrast to absent ones when evaluating which of two alternative hypotheses provides a better account for a set of observations. We also explored whether two features, namely the ratio of present-to-absent features in each set of observations and the presentation format of the features' probabilities (i.e., by making explicit either the probabilities of occurrences, non-occurrences or both) can affect the tendency to overestimate the importance of present features. Finally, we explored correlationally whether the informative strength of the set of present or absent features can affect that tendency.

Design

The quantitative parameters and formal properties denoting each one of the problems that we used are provided in Appendices I and J. We presented to a total of 126 participants a series of 18 abstract problems, each one describing an array of 4 to 5 features whose probability distributions under the two alternative hypotheses were fully explicit. In each problem, some features were present (either two or three) and the others (either three or two) absent. In all problems, the two subsets of present and absent features pointed in opposite directions: Namely, if the present features taken alone supported hypothesis 1, then the absent features supported hypothesis 2, and vice versa. The 18 problems were devised according to a 3 × 3 fully within-subjects design (with two different problem versions in each cell), factoring the type of the correct response and the ratio of present-to-absent features. The correct response—namely the hypothesis most probable if taking into account all of the features, including present and absent ones—could match either the hypothesis suggested by the present features alone (labeled "presence-consistent" problems), the hypothesis suggested by the absent features alone ("absence-consistent" problems) or none of the above ("equiprobable" problems, in which the pattern of features was equally likely under the two alternative hypotheses). We varied the ratio of

present-to-absent features orthogonally to the previous factor, because it might affect either the occurrence or the strength of feature-positive effects. Indeed, if it is true that feature-positive effects descend from the fact that, in general, occurrences are less likely than non-occurrences (e.g., Newman et al., 1980), then scenarios in which the number of absent features are less than the number of present features could direct attention to the former and improve the chances that they are attended. Therefore, in six of the problems, present and absent features came in the same number (ratio of present-to-absent = 2:2); in six other problems, there were more present than absent features (3:2), and in the remaining six problems, there were less present than absent features (2:3) (this manipulation also varied the overall amount of features, four in some problems and five in others).

Embedded within the main factorial design described above, we also varied non-orthogonally the informational strength of the sets of features in order to allow correlational analyses between the informational strength and participants' choices. In the 12 non-equiprobable problems, the 4 or 5 clues overall conveyed .23 bits of information [corresponding to an increase in the probability of the correct hypothesis from the initial p (H) = .5 to p (H | E) = .78]. This value is above the average threshold of information sensitivity that was measured in 130 non-expert participants engaged in abstract tasks similar to the present ones in three previous studies (reported in Cherubini et al., 2009), which was between .12 and .18 bits. In the equiprobable problems, of course, the whole set of clues overall conveyed 0 bits of information.

In non-equiprobable problems, the two subsets of present and absent features conveyed (if their ΔI was measured while ignoring the other set) from .92 to .98 bits of information each—that is, they were rather strong. Equiprobable problems were used for presenting weaker subsets of features so that the ΔI of the subsets of features was varied on eight levels (from very low to very high) for present features and on nine similar levels for absent features (see Appendices I and J).

Across the three experiments, we planned to control whether the format of the probability information affected the occurrence or the magnitude of feature-positive effects. In all previous hypothesis-testing studies that used explicit probabilities, values were used to describe the probabilities of feature occurrences. The complementary probabilities of non-occurrences, thus, were implicit and had to be derived by the participants. We conjectured that the explicit presentation of non-occurrence probabilities

might reduce the cognitive load required to take them into proper account (for a similar argument see Experiment 2, Chapter 3), and, at the same time, draw attention to their diagnostic relevance, thus possibly weakening feature-positive effects. In the first experiment, we only presented the probabilities of occurrences (the most typical manipulation used in previous studies). In the second experiment, we presented both the probabilities of occurrences and the complementary probabilities of non-occurrences. In the third experiment, we only presented the probabilities of non-occurrences.

Main dependent variables and main predictions

In all experiments, responses were primarily classified as presence-consistent, absence-consistent, or equiprobable. Presence-consistent responses were those mentioning the deck that was supported by the present features (regardless of whether they were correct responses or not), and similarly absence-consistent responses reflected choices for the deck supported by the absent features. According to this classification, a feature-positive effect should manifest itself as an increase of present-consistent responses with respect both to chance levels and to absent-consistent and equiprobable responses. For the sake of further analyses, responses were re-classified as correct or incorrect: Correct responses were those in which the deck supported by the whole set of features was chosen for non-equiprobable problems as well as those in which equiprobable responses were made in response to equiprobable problems; all the other responses were deemed incorrect. According to the latter classification, a feature-positive effect should manifest itself as an increase in correct responses for presence-consistent problems as compared to absence-consistent and equiprobable problems. Modulation of feature-positive effects by the present-to-absent features ratio or, between experiments, by the format of the probabilistic information is possible: Specifically, we expected that more attention should be apportioned to absent features in problems in which they are rare (3:2 problems) and in Experiments 2 and 3, in which the probabilities of non-occurrences are explicitly reported. Finally, we asked all participants to rate their confidence in each response on a 1-to-7 rating scale. According to this variable, a feature-positive effect might be observed by an increase in confidence when responses are presence-consistent rather than absence-consistent or equiprobable (or, in terms of correct/incorrect responses, by an increase of confidence in correct responses to presence-consistent problems as opposed to correct responses for all other problems).

Materials and procedure

In each one of the 18 problems, written instructions described two decks, deck 1 and deck 2, each made up of 100 cards. Each card within each deck featured from zero to five letters (zero to four letters for the six problems in which the ratio of present and absent features was 2:2), chosen from the set {B, C, D, F, G} (G was omitted from the contents of the problems in which the maximum number of letters was four). The instructions stated that the presence or absence of a letter on a card was unrelated to the presence or absence of any other letter and that each letter could be reported at most once on each card. A table fully illustrated the probabilities of the occurrence (Experiment 1), the occurrence and non-occurrence (Experiment 2), or the non-occurrence (Experiment 3) of each letter in each deck. Participants were told that the experimenter drew a card from a randomly selected deck (i.e., the prior probability of each deck was .5). The content of the drawn card was described and pictorially shown to participants. Specifically, they were told the letters (either two or three) depicted on the card and those that were absent. Given this information, participants were asked to determine from which deck the card was most likely drawn. Sample problems for the three experiments are provided in Appendix K. The probability parameters and formal properties of each one of the 18 problems are reported in Appendices I and J.

The order of presentation for the three alternative conclusions in each problem (i.e., "equiprobable", "deck 1" or "deck 2") was fully balanced across participants so that six versions of the questionnaire were created. Although in the appendices the present letters are always reported as B and C or B, C and D, the actual letters which were present or absent in each version of the problem were randomized across problems. Each problem was printed on a separate sheet of paper, and booklets were prepared presenting them in random order. Participants were individually approached in libraries and study rooms at the University of Milano-Bicocca. They were asked to participate in a study on the hypothesis-testing process, and those who accepted were given the experimental booklet. On the cover page, some personal data (e.g., age, gender, and years of education) were collected. In order to familiarize participants with the task, the second page provided the instructions and a sample problem with detailed explanations about the task and its requirements. Upon completion of each problem, participants were asked to express their confidence on the correctness of their answers on a 7-point scale.

Experiment 1

In this first experiment, we tested whether feature-positive effects can be observed and whether their magnitude is modulated by the ratio of present-to-absent features and by the informativeness of the set of present clues in tasks in which only the probabilities of features' occurrence are made explicit to the participants (Appendix K, Table K.1). This is the most common probabilistic format found in previous hypothesis-testing research.

Method

Design, procedure and materials

As *per* the general design and procedure described in the overview of the studies paragraph.

Participants

A total of 42 graduate and undergraduate students (18 females, 24 males; mean age = 22.7 years, range: 20-29 years; mean education = 16 years, *SD* = 1.7) of the University of Milano-Bicocca volunteered to take part in the experiment.

Results

Comparisons with chance level

Table 4.1 reports the mean number and percentage of presence-consistent, absence-consistent and equiprobable responses for each one of the nine experimental cells derived by the type of response × ratio of present-to-absent features experimental design. There were two problems in each experimental cell, and thus the mean number of responses ranged from 0 to 2. Percentages were computed out of 84 total responses (because of rounding, some row totals do not exactly equal 2 for means and 100 for percentages). Correct responses are in bold. The asymptotic *p* of the binomial tests comparing actual answers to a chance level of 33% are reported as "*", meaning $p < .05$, "**", meaning $p < .01$, or "***", meaning $p < .001$.

In all conditions but one, presence-consistent responses were significantly more frequent than chance. The exception was the condition in which the two decks were equiprobable and the ratio of present-to-absent features was 3:2. Again, with the exception of that condition, the frequencies of equiprobable responses were significantly

less than chance in all conditions. The absence-consistent responses were at chance level in most conditions. They dropped below chance level in the condition in which the correct response was presence-consistent and the ratio was either 2:2 or 2:3 as well as the condition in which the two decks were equiprobable and the ratio was 2:3. These findings hint at a strong feature-positive effect. The response suggested by the present features was the preferred one in most conditions, both when it was the correct response (upper row in Table 4.1) and when it was incorrect (middle and bottom row in Table 4.1), corroborating the conjecture that present features are the ones most considered when evaluating which hypothesis fits best with a set of data.

		Responses:		
	Present-to-absent ratio	*Presence-consistent*	*Absence-consistent*	*Equiprobable*
Presence	*2:2*	**1.5 (75%) *****	.33 (17%) ***	.17 (8%) ***
consistent	*3:2*	**1.12 (56%) *****	.64 (32%)	.24 (12%) ***
Problems	*2:3*	**1.31 (65%) *****	.40 (20%) **	.29 (15%) ***
Absence	*2:2*	1.14 (57%) ***	**.60 (30%)**	.26 (13%) ***
consistent	*3:2*	1.02 (51%) ***	**.76 (38%)**	.21 (11%) ***
Problems	*2:3*	1.19 (60%) ***	**.60 (30%)**	.21 (11%) ***
Equiprobable problems	*2:2*	.90 (45%) **	.62 (31%)	**.48 (24%) ***
	3:2	.76 (38%)	.55 (27%)	**.69 (35%)**
	2:3	1.12 (56%) ***	.40 (20%) **	**.48 (.24%) ****

Table 4.1. Mean number (ranging from 0 to 2) and percentage of each type of choice in each type of problem in Experiment 1. There were 18 problems (2 of each type), $N = 42$. The stars report the level of significance against chance level (set at .33): * = $p < .05$; ** = $p < .01$; *** = $p < .001$. Correct responses are in bold.

Correct responses and presence-consistent responses

Table 4.1 hints at a possible interaction between the type-of-correct-response factor and the present-to-absent-features-ratio factor. In order to explore this interaction, we analyzed the mean rates of correct responses (the bold diagonal in Table 4.1). Because the

ANOVA is an improper test for count data ranging from zero to two *per* cell (e.g., Jaeger, 2008), we ran a generalized linear model for repeated measures model with a Poisson distribution for the response variable by means of the SAS™ statistical package, factoring the type of correct response (presence-consistent vs. absence-consistent vs. equiprobable) and the ratio of present-to-absent features (2:2 vs. 3:2 vs.2:3). The main effect of the type of correct response was significant, $\chi^2 = 24.44$, df = 2, $p < .0001$ ($M_{\text{presence-consistent}} = 1.31$, $M_{\text{absence-consistent}} = .65$, $M_{\text{equiprobable}} = .55$), confirming that correct responses were more frequent in the presence-consistent than in the absence-consistent, pair-wise comparison, Bonferroni correction: χ^2=11.59, df=1, p=.0007, or equiprobable, $\chi^2 = 24.31$, df = 1, $p <$ .0001, conditions. The main effect of the ratio of present-to-absent features was not significant, $\chi^2 = 2.47$, df = 2, $p = .29$. Beyond suggesting that the ratio of present-to-absent features does not have by itself a main influence on the frequency of correct responses, this finding also shows that the different number of clues in the three conditions (five clues in the 3:2 and 2:3 conditions vs. four clues in the 2:2 conditions) did not have appreciable effects on responses. The two-way interaction was significant, $\chi^2 = 11.29$, df = 4, $p < .05$. The interaction probably emerged from the decrease of correct responses for presence-consistent problems in the 3:2 present-to-absent ratio condition and from the increase of correct responses in the absence-consistent and equiprobable problems in the 3:2 present-to-absent ratio condition (see Table 4.1, bold diagonal). This finding is consistent with the prediction that absent clues are apportioned more attention when they are less frequent than present clues. A similar trend, this time indicated by a main effect for the ratio of present-to-absent features, was observed for the occurrence of presence-consistent responses, regardless of their correctness (column 1 of Table 4.1). We statistically explored this interaction by means of another generalized Poisson model, featuring the number of presence-consistent responses as the dependent variable and factoring the type of problem and the ratio of present-to-absent features. The analysis yielded a significant main effect for the ratio of present-to-absent features, $\chi^2 = 8.65$, df = 2, $p = .0132$ ($M_{2:2\ \text{problems}} = 1.18$; $M_{3:2\ \text{problems}} = .97$; $M_{2:3\ \text{problems}} = 1.21$). This finding corroborates the idea that present features drive attention less when the absent features are rare than when they are equally frequent or more frequent than present features. The main effect of the type-of-problems factor was also significant, χ^2=14.18, df=2, p<.001 ($M_{\text{presence-consistent problems}} = 1.31$, $M_{\text{absence-consistent problems}} = 1.12$, $M_{\text{equiprobable problems}} = .93$), suggesting that, although participants in aggregate form had an overall preference for

presence-consistent responses, they were also sensitive to the formal correctness of the response, presence-consistent vs. equiprobable: $\chi^2 = 12.43$, df = 1, $p = .0004$; presence-consistent vs. absence-consistent: $\chi^2 = 5.58$, df = 1, $p = .0181$; absent-consistent versus equiprobable was not significant, $\chi^2 = 3.15$, df = 1, $p = .0757$. The two-way interaction was not significant.

Confidence ratings

Mean confidence ratings for the correctness of the response, derived from a 1-to-7 rating scale, in which 7 indicates extreme confidence and 1 reflects no confidence, was 4.73 for presence-consistent responses, 4.36 for absent-consistent responses, and 4.13 for equiprobable responses. In keeping with our hypotheses, two 1-tailed exact Wilcoxon tests showed that confidence toward present-consistent responses was significantly higher than confidence toward absence-consistent responses, $N = 37$ (five participants lacked absence-consistent responses), $Z = 2.01$, $p = .022$, and confidence toward equiprobable responses, $N = 35$ (seven participants lacked equiprobable responses), $Z = 3.48$, $p = .0001$. Because of lack of a directional hypothesis, the comparison between absence-consistent and equiprobable responses was two-tailed, $N = 31$ (11 participants lacked either equiprobable or absence-consistent responses or both), yielding a non-significant difference, $Z = 1.85$, $p = .07$. As for confidence toward correct responses, there were too many missing data in the full factorial design type of correct response × ratio of present-to-absent features to compute the analysis. Accordingly, we pooled together the data across different levels of the present-to-absent ratio factor and compared confidence ratings toward correct responses to the presence-consistent problems ($M = 4.84$) with those for the absence-consistent problems ($M = 4.33$) and equiprobable problems ($M = 4.13$). Consistent with our expectations, two 1-tailed exact Wilcoxon tests showed that confidence towards correct responses was higher for presence-consistent problems than for either absence-consistent problems, $N = 30$, $Z = 1.66$, $p = .049$, or equiprobable problems, $N = 32$, $Z = 3.03$, $p = .001$. The comparison between confidence in correct responses for absence-consistent and equiprobable problems was two-tailed and was not significant, $N = 25$, $Z = 1.7$, $p = .085$.

Finally, mean confidence across the problems was positively correlated with the number of presence-consistent responses chosen by participants, Spearman's rho = .56, $N = 18$, $p < .05$, two-tailed. It did not correlate significantly with the number of correct

responses, rho = −.21, p = .4, nor with the number of absence-consistent responses, r = −.1, p = .68.

Correlations considering the ΔI of different subsets of clues

By aggregating data across participants for the 18 problems, we investigated the correlations between the number of presence-consistent and absence-consistent choices in each problem and the ΔI in bits conveyed in that problem by the two subsets of present or absent clues (see Appendices I and J for the exact values). The ΔI of present clues was strongly correlated with the number of presence-consistent choices, rho = .73, N = 18, p < .001, two-tailed, meaning that, in aggregated form, participants were sensitive to the formal amount of information conveyed by the present clues. That is, the more the present clues were informative, the more likely the choice of the presence-consistent response was. Conversely, no similar correlation occurred between absence-consistent choices and the ΔI of absent clues, rho = .19, N = 18, p = .43, meaning that the amount of information conveyed by absent clues did not appreciably affect the decision to choose or not choose the absence-consistent response. These two findings further clarify our previous results by suggesting that participants were only or mostly sensitive to the formal amount of information for present features but not for absent ones.

Discussion

The results of Experiment 1 provide evidence that, when evaluating competing hypotheses, people pay more attention to occurring features than to non-occurring ones. In all but one condition, the participants chose the hypothesis indicated by the present features (which, in all problems, was opposite to the one indicated by the absent features) above chance level. Preferences for other responses were always either at chance level or significantly below it. In addition, the rate of correct responses was significantly higher when those responses were suggested by present features than in all other cases. Confidence toward the correctness of one's own judgments was significantly higher for presence-consistent responses than for absence-consistent or equiprobable responses, and it was significantly higher for correct responses in accordance with present clues than for any other correct response. Furthermore, mean confidence ratings across problems correlated positively and significantly with the number of presence-consistent responses. The feature-positive effect did not completely cancel out the participants' sensitivity to

formally correct responses, as suggested by the finding that presence-consistent responses were significantly more frequent when they were also correct responses. Nevertheless, sensitivity to the formal amount of information conveyed by the stimuli was significant only for present clues: The formal amount of information (ΔI) conveyed by those clues strongly correlated with the number of presence-consistent choices in that problem across participants, whereas the ΔI conveyed by the subset of absent clues did not correlate significantly with the number of absence-consistent choices.

As an ancillary result, the feature-positive effect in hypothesis evaluation is apparently modulated by the rate of present-to-absent features. In this experiment, the effect was weaker when the number of absent features was less than the number of present features (2 vs. 3, respectively). This trend is possibly brought about by a "rarity" effect, namely the tendency to focus attention and apportion more weight to unusual events than to usual ones (McKenzie 2004; McKenzie & Mikkelsen, 2000, 2007). In this instance, the rarity of the absent features in the 3:2 present-to-absent features ratio problems partly counterbalanced the tendency to predominantly pay attention to present features.

Experiment 2

Providing participants with fully explicit information about the probabilities of occurrences only, as *per* the procedure of Experiment 1, might contribute to the observed tendency to attend to absent features less than present ones, because, whereas the probability of the occurrence of each feature was readily available in table format, the probability of its non-occurrence had to be inferred by complementation. In this experiment, we attempted to reduce the computational effort associated with the processing of the absent clues by providing participants with explicit information, not only about the probability of the presence of each clue, but also about the probability of its absence (Appendix K, Table K.2).

Method

Participants

A total of 42 volunteers (23 females, 19 males; mean age = 22.6 years, range: 19-32 years; mean education = 15.8 years, *SD* = 1.5) took part in the experiment. None had taken part in Experiment 1.

Materials and procedure

The materials and design were identical to those used in Experiment 1, except for the presentation format. Indeed, as in Experiment 1, participants were told the number of cards reporting each letter within each deck and, in addition, the number of cards that did not report a certain letter within each deck. An example of the presentation format in this experiment is reported in Appendix K, Table K.2.

Results

Comparisons with chance level

Table 4.2 reports the mean number and percentage of presence-consistent, absence-consistent and equiprobable responses in each experimental condition. Correct responses are in bold. The asymptotic p of the binomial tests comparing actual answers to a chance level of 33% are reported as "*", meaning $p < .05$, "**", meaning $p < .01$, or "***", meaning $p < .001$.

		Responses:		
	Present-to-absent ratio	*Presence-consistent*	*Absence-consistent*	*Equiprobable*
Presence consistent Problems	*2:2*	**1.36 (68%) *****	.31 (15%) ***	.33 (17%) ***
	3:2	**1.38 (69%) *****	.26 (13%) ***	.36 (18%) **
	2:3	**1.33 (67%) *****	.40 (20%) **	.26 (13%) ***
Absence consistent Problems	*2:2*	1.05 (52%) ***	**.45 (23%) ***	.5 (25%)
	3:2	1.26 (63%) ***	**.43 (21%) ***	.31 (15%) ***
	2:3	1.23 (62%) ***	**.55 (27%)**	.21 (11%) ***
Equiprobable problems	*2:2*	.86 (43%) *	.43 (21%) *	**.71 (36%)**
	3:2	.48 (24%) *	.93 (46%) **	**.6 (30%)**
	2:3	1.07 (54%) ***	.64 (32%)	**.29 (14%) *****

Table 4.2. Mean number (ranging from 0 to 2) and percentage of each type of choice in each type of problem in Experiment 2. There were 18 problems (2 per cell), $N = 42$. The stars report the level of significance against chance level (set at .33): * = $p < .05$; ** = $p < .01$; *** = $p < .001$. Correct responses are in bold.

In all conditions but one, presence-consistent responses were significantly more frequent than chance. The exception was the same as in Experiment 1—that is, the condition in which the two decks were equiprobable and the ratio of present-to-absent features was 3:2. In that condition, presence-consistent responses were significantly below chance level (in Experiment 1, they were at chance level). Equiprobable responses were at chance level or significantly below it in all conditions. The absence-consistent responses were below chance level in most conditions, except for equiprobable 3:2 problems (in which they were above the chance level of 33%; this was the only condition in which they were preferred to presence-consistent responses) and equiprobable 2:3 problems (in which they were at chance level). These preliminary tests apparently replicated the strong feature-positive effect observed in the previous experiment: The response suggested by the present features was the preferred one in most conditions, both when it was the correct response (upper row in Table 4.2) and when it was incorrect (middle and bottom row in Table 4.2), with only one exception.

Correct responses and presence-consistent responses

The frequency of correct responses was analyzed by means of a generalized repeated-measures model for a Poisson distribution, factoring the type of correct response (presence-consistent vs. absence-consistent vs. equiprobable) and the ratio of present-to-absent features (2:2 vs. 3:2 vs. 2:3). The main effect for the type of correct response was significant, $\chi^2 = 19.98$, df = 2, $p < .0001$ ($M_{\text{presence-consistent}} = 1.36$, $M_{\text{absence-consistent}} = .48$, $M_{\text{equiprobable}} = .53$), confirming that correct responses were more frequent in presence-consistent than in the absence-consistent, pair-wise comparison, Bonferroni correction: $\chi^2 = 19.17$, df = 1, $p < .0001$, or equiprobable, $\chi^2 = 24.99$, df = 1, $p < .0001$, conditions. The main effect of the ratio of present-to-absent features was not significant, $\chi^2 = 4.42$, df = 2. The two-way interaction was significant, $\chi^2 = 10.9$, df = 4, $p < .05$. The statistical results closely match those of Experiment 1. However, in Experiment 1, the interaction was driven by a relative increase in correct responses for absence-consistent and equiprobable problems with a 3:2 ratio of present-to-absent features along with a decrease in accuracy in the 3:2 presence-consistent problems. By contrast, in this experiment, the interaction was probably driven by the relative increase of correct responses for the 2:2 and 3:2 equiprobable problems only (see Table 4.2, bold diagonal). Similar to Experiment 1, a generalized Poisson model featuring the number of presence-consistent responses as the

dependent variable and factoring the type of problem and the ratio of present-to-absent features yielded significant main effects for the type of problem, $\chi^2 = 19.98$, df = 2, $p < .0001$ ($M_{\text{presence-consistent problems}} = 1.36$, $M_{\text{absence-consistent problems}} = 1.18$, $M_{\text{equiprobable problems}} = 0.80$) and the ratio of present-to-absent features, $\chi^2 = 8.17$, df = 2, $p < .05$ ($M_{2:2\text{ problems}} = 1.09$, $M_{3:2\text{ problems}} = 1.04$, $M_{2:3\text{ problems}} = 1.21$). The former effect replicates the one found in Experiment 1, showing that, beyond generally preferring the presence-consistent responses, participants were also partly sensitive to the correct responses. The latter effect shows a tendency for the preference towards present-consistent responses to decrease for the problem versions in which there were two absent clues (i.e., the 2:2 and 3:2 problems) as compared to those in which they were three (the 2:3 problem versions). The two-way interaction was also significant, albeit in a different manner from Experiment 1, $\chi^2 = 14.28$, df = 4, $p < .01$ (see the first column of Table 4.2). The interaction is difficult to interpret, inasmuch as it probably emerged from the increase in present-consistent responses in the 2:3 condition of the equiprobable problems in comparison to the 2:2 and 3:2 conditions. It might derive from a rarity effect, this time favoring present features, but this sort of effect was not observed in the previous experiment and is weak in the absence-consistent problems.

Confidence ratings

The mean confidence ratings for the correctness of the responses were 4.77 for presence-consistent responses, 4.49 for absent-consistent responses, and 3.93 for equiprobable responses. Two 1-tailed exact Wilcoxon tests showed that confidence toward present-consistent responses was significantly higher than confidence toward equiprobable responses, $N = 33$, $Z = 3.07$, $p = .001$, but—differently from Experiment 1 and from predictions—it was not significantly higher than confidence toward absence-consistent responses, $N = 39$, $Z = .62$, $p = .33$. Confidence in absence-consistent responses was significantly higher than confidence toward equiprobable responses, $N = 30$, $Z = 2.48$, $p = .015$, two-tailed. As for confidence toward correct responses, we pooled together the data across different levels of the present-to-absent ratio factor, and compared confidence ratings toward correct responses in the presence-consistent problems ($M = 5$) with those in absence-consistent problems ($M = 4.38$) and equiprobable problems ($M = 3.87$). Two 1-tailed exact Wilcoxon tests showed that, in keeping with our expectations, confidence towards correct responses was higher in presence-consistent problems than in absence-

consistent problems, $N = 22$, $Z = 1.98$, $p = .032$, and equiprobable problems, $N = 29$, $Z = 2.69$, $p = .004$. Differences in confidence between correct responses in absence-consistent problems and correct responses in equiprobable problems were not significant, $N = 17$, $Z = .96$, $p = .37$, two-tailed. Finally, replicating a similar result observed in Experiment 1, mean confidence across the problems was positively correlated with the number of presence-consistent responses chosen by participants, Spearman's rho = .55, $N = 18$, $p < .05$, two-tailed. It did not correlate significantly with either the number of correct responses, rho = .12, $p = .65$, or with the number of absence-consistent responses, $r = -.11$, $p = .66$.

Correlations considering the ΔI of different subsets of clues

Similarly to what occurred in Experiment 1, the ΔI in bits conveyed by present clues in each problem was strongly correlated with the number of presence-consistent choices on that problem, aggregated across participants, rho = .86, $N = 18$, $p < .0001$, two-tailed, but the ΔI of the subsets of absent features did not correlate with the number of absence-consistent choices, rho = −.25, $N = 18$, $p = .32$. Again, this finding hints at the fact that participants, in aggregated form, were sensitive to the formal amount of information conveyed by the present clues but not to information conveyed by absent clues.

Discussion

In Experiment 2, participants received explicit probabilistic information concerning not only the probabilities of the presence of each feature under the two hypotheses but also the complementary probabilities of their absence. This manipulation did not strongly weaken the feature-positive effect observed in Experiment 1, as shown by many results that closely matched that experiment. In all conditions but one, the participants' choices of the hypothesis supported by the present features were above chance level while preferences for other responses were either at chance level or significantly below it (with one exception, namely the choice of absence-consistent responses in the 3:2 versions of the equiprobable problems). The rate of presence-consistent correct responses was higher than the rates of absence-consistent or equiprobable correct responses. Confidence regarding one's own judgments was higher for presence-consistent correct responses than for any other correct responses, and the mean confidence ratings across problems correlated positively and significantly with the number of presence-consistent responses.

Finally, also in this experiment, the feature-positive effect did not completely cancel out participants' sensitivity to formally correct responses, as shown by the finding that presence-consistent responses were significantly more frequent when they were correct, but sensitivity to the formal amount of information conveyed by the stimuli was appreciable for present clues only.

However, in some other respects, the results differed from those of Experiment 1. Confidence toward presence-consistent responses was not significantly higher than confidence toward absence-consistent responses, hinting at a marginal weakening of the feature-positive effect as far as confidence ratings were concerned. The 3:2 present-to-absent ratio versions of the problems did not uniformly weaken the preference for present-consistent responses, as occurred in Experiment 1. In this experiment, the ratio of present-to-absent features had different and apparently non-systematic effects: Presence-consistent responses were more frequent in the 2:3 equiprobable problems than in the 3:2 and 2:2 versions, and absent-consistent responses were chosen significantly more than chance in the 3:2 version of the equiprobable problems. In some, but not all, instances, 3:2 problems apparently drove attention to absent features, and 2:3 versions apparently drove attention to present features. The pattern suggests that making the probabilities of occurrences and non-occurrences of the features explicit had a heterogeneous impact on the effects of the rarity of those features.

In conclusion, the results of Experiment 2 corroborate the main findings of Experiment 1, as far as the choice of responses is concerned. Explicating the probabilities of non-occurrences affected only the confidence toward responses and the effects of the ratio of present-to-absent features.

Experiment 3

In Experiment 2, the explication of the probability of non-occurrences did not appreciably weaken the tendency observed in Experiment 1 to over-attend to present features in hypothesis testing, at least as far as response selection was concerned. Participants still preferred to select the hypothesis supported by present features, which was opposite to the one supported by absent features. Apparently, the overt presentation of the probabilities of occurrences had relatively minor effects, by making less systematic the effect of the present-to-absent feature ratio and by reducing somewhat the participants' confidence toward presence-consistent responses. In Experiment 3, we exclusively presented to

participants the probabilities of non-occurrences. That is, in this experiment, it was the consideration of present features that required one more cognitive step, namely the complementation of the probabilities of absence in order to derive the probabilities of presence. In this respect, Experiment 3 was symmetrically opposite of Experiment 1, and the goal was to enhance the consideration of absent features. In all other respects, the stimuli, design, procedure and methods were exactly the same as those used in the two previous experiments. An example of the format of the problems presented in Experiment 3 is displayed in Appendix K, Table K.3.

Method

Participants

A total of 42 volunteers (21 females, 21 males, mean age = 22.2 years, range: 20-27 years; mean education = 15.7 years, *SD* = 1.7) took part in the experiment. None had taken part in the previous two experiments.

Results

Comparisons with chance level

Table 4.3 reports the mean number and percentage of presence-consistent, absence-consistent and equiprobable responses in each experimental condition. Correct responses are in bold. The asymptotic *p* of the binomial tests comparing actual answers to a chance level of 33% are reported as "*", meaning $p < .05$, "**", meaning $p < .01$, or "***", meaning $p < .001$.

In all of the presence-consistent and absence-consistent problems, presence-consistent responses were significantly more frequent than chance. An exception was with equiprobable problems, with presence-consistent responses at chance level in all conditions, whereas, in Experiments 1 and 2, they were at or below chance levels only in the 3:2 versions of the equiprobable problems. Absence-consistent responses, which were mostly below chance levels in the previous experiments, were mostly at chance level in the present experiment, possibly indicating a marginal increase in the attendance to absent features. Equiprobable responses were mostly below chance level, except for the equiprobable problems, in which they were at chance level. Divergences from Experiments 1 and 2 are small: the overall pattern still suggests a rather strong, quite generalized preference for attending to present features over absent ones.

		Responses:		
	Present-to-absent ratio	*Presence-consistent*	*Absence-consistent*	*Equiprobable*
Presence	*2:2*	**1.31 (65%) *****	.52 (26%)	.17 (8%) ***
consistent	*3:2*	**1.24 (62%) *****	.36 (18%) **	.40 (20%) **
Problems	*2:3*	**1.05 (52%) *****	.52 (26%)	.43 (21%) **
Absence	*2:2*	1.05 (52%) *******	**.5 (25%)**	.45 (23%) *
consistent	*3:2*	1.21 (61%) *******	**.52 (26%)**	.26 (13%) ***
Problems	*2:3*	1.14 (57%) *******	**.5 (25%)**	.36 (18%) **
Equiprobable problems	*2:2*	.69 (34%)	.76 (38%)	**.55 (27%)**
	3:2	.55 (27%)	.74 (37%)	**.71 (36%)**
	2:3	.76 (38%)	.48 (24%) *	**.76 (38%)**

Table 4.3. Mean number (ranging from 0 to 2) and percentage of each type of choice in each type of problem in Experiment 3. There were 18 problems (2 per cell), $N = 42$. The stars report the level of significance against chance level (set at .33): * = $p < .05$; ** = $p < .01$; *** = $p < .001$. Correct responses are in bold.

Correct responses and presence-consistent responses

The rate of correct answers was analyzed by means of a generalized repeated-measures model for a Poisson distribution, factoring the type of correct response (presence-consistent, absence-consistent, equiprobable) and the ratio of present-to-absent features (2:2, 3:2, 2:3). The main effect for the type of correct response was once again significant, $\chi^2 = 11.14$, df = 2, $p < .005$ ($M_{\text{presence-consistent}} = 1.2$, $M_{\text{absence-consistent}} = .51$, $M_{\text{equiprobable}} = .67$), confirming that correct responses were more frequent in the presence-consistent than in the absence-consistent, pair-wise comparisons, Bonferroni correction: $\chi^2 = 10.59$, df = 1, $p = .0011$, or equiprobable, $\chi^2 = 9.15$, df = 1, $p = .0025$, problems. As occurred in the previous experiments, the main effect for the ratio of present-to-absent features was not significant, $\chi^2 = .80$, df = 2. However, in contrast to the previous experiments, the two-way interaction was also not significant: Hence, the ratio of present-to-absent features in

this version of the task had no appreciable effects whatsoever on the frequency of correct responses.

A second generalized Poisson model featured the number of presence-consistent responses as the dependent variable and factored the type of problem and the ratio of present-to-absent features. Similar to the two previous experiments, it yielded a significant main effect for the type of problem, $\chi^2 = 17.58$, df = 2, $p < .0005$ ($M_{\text{presence-consistent problems}} = 1.2$, $M_{\text{absence-consistent problems}} = 1.13$, $M_{\text{equiprobable problems}} = .67$), but surprisingly the effect shows that the presence-consistent responses were more frequent in the presence-consistent problems in comparison to the equiprobable problems, pair-wise comparison, Bonferroni correction: $\chi^2 = 16.64$, df = 1, $p < .0001$, and in the absence-consistent problems in comparison to the equiprobable problems, $\chi^2 = 146.96$, df = 1, $p < .0001$. Presence-consistent responses were not significantly more frequent in presence-consistent vs. absence-consistent problems, $\chi^2 = .60$, df = 1, $p = .4368$. If anything, this pattern hints at a strengthening, instead of a weakening, of the feature-positive effect in this version of the task, as far as the rate of presence-consistent responses are concerned. The main effect of the ratio of present-to-absent features was not significant, $\chi^2 = .27$, df = 2, nor it was the two-way interaction, $\chi^2 = 6.41$, df = 4, confirming that the rarity of features in this experiment did not appreciably affect the overall preference for the responses suggested by present features.

Confidence ratings

The mean confidence ratings for the correctness of the responses were 4.4 for presence-consistent responses, 4.36 for absent-consistent responses, and 3.95 for equiprobable responses. Two 1-tailed exact Wilcoxon tests showed that confidence toward presence-consistent responses was significantly higher than confidence toward equiprobable responses, $N = 36$, $Z = 2.58$, $p = .004$, but, contrary to both Experiment 1 and our initial expectations, albeit in keeping with the results of Experiment 2, it was not significantly higher than confidence toward absence-consistent responses, $N = 38$, $Z = 1.35$, $p = .09$. Confidence toward absence-consistent responses was significantly higher than confidence toward equiprobable responses, $N = 37$, $Z = 2.06$, $p = .04$, two-tailed, which was similar to the results observed in Experiment 2. A similar pattern was observed for confidence ratings toward correct responses: There were no reliable differences between presence-consistent ($M = 4.5$) and absence-consistent responses ($M = 4.36$), $N = 18$, $Z = .05$, $p =$

.49, one-tailed, whereas there were differences between presence-consistent and equiprobable (M = 3.71) responses, N = 29, Z = 2.9, p = .001, one-tailed, and a trend toward a difference between absence-consistent and equiprobable responses, N = 19, Z = 1.93, p = .054, two-tailed, p = .027, one-tailed; one-tailed testing could be justified here by the corresponding significant difference observed in Experiment 2. These findings suggest that, although the overt presentation of the probabilities of non-occurrences alone did not remarkably weaken the feature-positive effect with regards to the participants' choices, it decreased the participants' confidence that judgments mostly based on present features were sound. However, as occurred in both previous experiments, mean confidence across the problems was also positively correlated with the number of presence-consistent responses in Experiment 3, Spearman's rho = .72, N = 18, p < .001, two-tailed, whereas it did not correlate significantly with either the number of correct responses, rho = .24, p = .33, or with the number of absence-consistent responses, rho = −.16, p = .53.

Correlations considering the ΔI of different subsets of clues

Similarly to what occurred in Experiments 1 and 2, the ΔI conveyed by present clues in each problem was positively correlated with the number of presence-consistent choices on that problem aggregated across participants, rho = .76, N = 18, p < .0001, two-tailed, but the ΔI of the subsets of absent features did not correlate with the number of absence-consistent choices, rho = −.38, N = 18, p = .12. This finding further substantiates the idea that the participants were sensitive to the formal amount of information conveyed by the present clues but were less sensitive to the information conveyed by absent clues.

Discussion

The presentation format of the probabilistic information in Experiment 3 was symmetrically opposite of that used in Experiment 1. Whereas only the probabilities of the occurrence of each feature under each hypothesis were overtly communicated to the participants in Experiment 1, in Experiment 3 participants were exclusively informed of the probabilities of non-occurrences. If the apportionment of excessive attention to present features were mostly caused by the common practice of communicating only the probabilities of occurrences, then in Experiment 3 the feature-positive effect should have disappeared or have been greatly weakened. By contrast, even in this condition, a

remarkably strong feature-positive effect emerged, as illustrated by many converging findings. First of all, the participants' choices of the hypothesis indicated by the present features were above chance level in all conditions, except for equiprobable problems (in which they were at chance level). The rate of correct responses was above chance level only when they were consistent with the responses backed by the present features. Furthermore, correct responses were significantly more frequent in the presence-consistent problems than in all other problems. The confidence ratings across problems correlated positively and significantly with the number of presence-consistent responses but not with the number of correct or absent-consistent responses. Replicating the results of the previous experiments, participants were sensitive to the formal amount of information conveyed by present clues (as showed by correlations between the ΔI and the frequency of choices) but not to the information conveyed by absent clues. Finally, in this experiment, the presence-consistent responses in presence-consistent problems (which were correct) and in absence-consistent problems (which were not correct) did not differ significantly. This latter finding suggests that, at least in this respect, the feature-positive effect was stronger in this experiment than in the previous two, in which the presence-consistent responses were significantly more frequent in presence-consistent problems than in absence-consistent problems.

The explicit presentation of the probabilities of non-occurrences apparently had minor effects on choices: Namely, absence-consistent responses increased from below-chance level (in the previous experiments) to chance level, and presence-consistent responses decreased to chance level in equiprobable problems. Also, this version of the task apparently cancelled out all effects of the ratio of present-to-absent clues. Finally, partially replicating the results of Experiment 2, the overt presentation of the probabilities of non-occurrences affected confidence ratings: Mean confidence toward presence-consistent responses was not significantly higher than confidence toward absence-consistent responses, and mean confidence toward correct presence-consistent responses was not higher than mean confidence toward correct absence-consistent responses.

In conclusion, the main results of Experiment 3 show that a feature-positive effect still influences the evaluation stage of hypothesis development, even when the probabilities of the non-occurrences of each feature under each alternative hypothesis are the only data available to the participants.

Cross-experimental analyses and discussion

The three experiments were run sequentially, and participants were not assigned randomly to the three samples. Apart from that, the experiments were homogeneous: The participants came from the same pool; the procedure and stimuli were the same, except for the presentation format of the probabilistic information; and the sample sizes were the same. Hence, a statistical cross-examination of the three experiments could theoretically be reliable. The mean number and percentage of choices (ranging from 0 to 2) for each response in each problem across the 126 participants (252 responses) to the three experiments are reported in Table 4.4. Correct responses are in bold. The asymptotic *p* of the binomial tests comparing actual answers to a chance level of 33% are reported as "*", meaning $p < .05$, "**", meaning $p < .01$, or "***", meaning $p < .001$.

		Responses:		
	Present-to-absent ratio	*Presence-consistent*	*Absence-consistent*	*Equiprobable*
Presence	*2:2*	**1.39 (70%) *** **	.39 (19%) ***	.22 (11%) ***
consistent	*3:2*	**1.25 (62%) *** **	.42 (21%) ***	.33 (17%) ***
Problems	*2:3*	**1.23 (62%) *** **	.44 (22%)***	.33 (17%) ***
Absence	*2:2*	1.08 (54%) ****	****.52 (26%)****	.40 (20%) ***
consistent	*3:2*	1.16 (58%) ****	****.57 (29%)****	.26 (13%) ***
Problems	*2:3*	1.19 (59%) ****	****.55 (28%)****	.26 (13%) ***
Equiprobable problems	*2:2*	.82 (41%)**	.60 (30%)	****.58 (29%)****
	3:2	.60 (27%)	.74 (37%)	****.67 (33%)****
	2:3	.98 (49%)***	.51 (25%) **	****.51 (25%)****

Table 4.4. Mean number (ranging from 0 to 2) and percentage of each type of choice in each type of problem in the three experiments. There were 18 problems (2 per cell), $N = 126$. The stars report the level of significance against chance level (set at .33): * = $p < .05$; ** = $p < .01$; *** = $p < .001$. Correct responses are in bold.

Presence-consistent responses were more frequent than chance in all conditions for both correct and incorrect responses, with the exception of the equiprobable problems

with a 3:2 ratio of present-to-absent features, in which they were at chance level. All other responses (again, both correct and incorrect ones) were at chance level or below it. The pattern hints at a strong feature-positive effect, which is only slightly modulated by the correctness of responses and by the ratio of present-to-absent features.

Correct responses

The relative frequency of correct responses (Table 4.4, bold diagonal) was analyzed by means of a generalized mixed model for a Poisson distribution, factoring the type of problems and the ratio of present-to-absent features within-participants and the format of probabilistic information as a between-groups variable. The main effect for the type of problems was significant, $\chi^2 = 55.87$, df = 2, $p < .0001$ ($M_{\text{presence-consistent problems}} = 1.29$, $M_{\text{absence-consistent problems}} = .55$, $M_{\text{equiprobable problems}} = .59$). This finding indicates that, independent of all other factors, namely the ratio of present-to-absent features and the presentation format of probabilistic information, correct responses were more frequent when they were backed by present features than when they were congruent with absent features, pair-wise comparison, Bonferroni correction: $\chi^2 = 40.\ 70$, df = 1, $p < .0001$, or were inconsistent with both present and absent features (i.e., equiprobable problems), $\chi^2 = 54.79$, df = 1, $p < .0001$. Notice that, because of the structure of the problems, this means that correct responses were more frequent when they were *opposite* to the response congruent with the absent features. The main effects for the ratio of present-to-absent features, $\chi^2 = 5.05$, df = 2, $p = .08$, and for the presentation format of the probabilistic information were not significant, $\chi^2 = 2.13$, df = 2, $p = .34$. All of the two-ways interactions did not reach significance. The three-way interaction was significant, $\chi^2 = 21.22$, df = 8, $p < .01$, probably originating from the different trends of the type of problems × ratio of present-to-absent features two-way interactions across the three experiments (see the individual discussions of each experiment).

Presence-consistent responses

A second generalized mixed model for a Poisson distribution with the same factors as the previous one was run to analyze the frequency of presence-consistent responses (Table 4.4, first column). It yielded a significant main effect for the type of problem, $\chi^2 = 49.46$, df = 2, $p < .0001$ ($M_{\text{presence-consistent problems}} = 1.29$, $M_{\text{absence-consistent problems}} = 1.14$, $M_{\text{equiprobable problems}} = .80$). Presence-consistent responses were significantly more frequent in presence-

consistent than absence-consistent, pair-wise comparison, Bonferroni correction: χ^2 = 8.46, df = 1, p = .0036, and equiprobable problems, χ^2 = 40.53, df = 1, p < .0001. However, presence-consistent responses were also significantly more frequent in absence-consistent than in equiprobable problems, χ^2 = 27.61, df = 1, p < .0001. Although the increased frequency of presence-consistent responses in presence-consistent problems, in which they were correct, in comparison to absence-consistent problems, in which they were incorrect, shows a residual sensitivity to the formal correctness of responses, their increased amount in comparison to equiprobable problems (that is apparent also for presence-consistent responses in absence-consistent problems) probably reflects the fact that, in the latter problems, the formal amount of information conveyed by present or absent clues was very small. Thus, the finding supports the idea that participants are mostly sensitive to the formal amount of information conveyed by present clues, as shown by correlations with the ΔI of the problems (see below). The main effect for the ratio of present-to-absent features was also significant, χ^2 = 12.62, df = 2, p < .005 ($M_{2:2\ \text{problems}}$ = 1.10, $M_{3:2\ \text{problems}}$ = 1.00, $M_{2:3\ \text{problems}}$ = 1.13). However, this effect is best accounted for by the significant type of problem × ratio of present-to-absent features two-way interaction, χ^2 = 22.96, df = 4, p < .0001 (means in the first column of Table 4.4), which shows that presence-consistent responses were indeed less frequent in the 3:2 problems, but only in the equiprobable problems. Thus, the rarity of absent clues can draw attention to absent features, albeit exclusively in circumstances in which the formal amount of information conveyed by the two subsets of present or absent clues is tiny. No other main effects or interactions reached significance.

Confidence ratings and sensitivity to ΔI across the three experiments

The mean confidence across the 18 problems in the three experiments was positively correlated with the number of presence-consistent responses, Spearman's rho = .71, N = 18, p < .001, two-tailed, whereas it did not correlate significantly with either the number of correct responses, rho = −.09, p = .73, or with the number of absence-consistent responses, rho = -.28, p = .26. That is, the more participants chose presence-consistent responses, the more they trusted their choices, whereas confidence did not appreciate as a function of either the actual number of formally correct choices or the number of absence-consistent choices.

The ΔI conveyed by present clues in each problem was positively correlated with the number of presence-consistent choices for that problem, rho = .89, N = 18, p < .0001, two-tailed, but the ΔI of the subsets of absent features did not correlate with the number of absence-consistent choices, rho = −.20, N = 18, p = .42. That is, participants were sensitive to the formal amount of information conveyed by the present clues, but they were not sensitive to the information conveyed by absent clues.

General discussion

These experiments lend conclusive support to one main finding and less strong support to some ancillary findings, which merit further investigation.

Main finding: A feature-positive effect influences the evaluation of alternative hypotheses

Many specific results of the three experiments and their synthesis, illustrated by the cross-experimental analyses, indicate univocally that people, when they evaluate available data for establishing which of two alternative and mutually exclusive hypotheses is the most likely, overrate the information conveyed by the occurrence of clues in comparison to that conveyed by the non-occurrence of other clues. Previous studies have reported this tendency (e.g., Fischhoff & Beyth-Marom, 1983; Slowiaczek et al., 1992), but no conclusive empirical evidence could directly support it. To our knowledge, the only study that directly investigated this issue with quasi-experimental methods failed to find support for it (Christensen-Szalansky & Bushyhead, 1981), although the authors attributed their negative finding to possible artifacts. In the present experiments, there are at least four sources of converging evidence for the occurrence of a rather strong feature-positive effect in the evaluation of alternative hypotheses:

1) In all experiments, the hypothesis consistent with the information conveyed by present clues and therefore inconsistent with the information conveyed by absent clues was preferred significantly above chance level in most conditions, regardless of whether it was the formally correct response or not. There were only a few exceptions, with presence-consistent responses at chance level, that emerged in some instances in which the two hypotheses were formally equiprobable. However, in those problems, the subset of present clues was formally very weak (that is, it conveyed a very low ΔI). Because participants

were sensitive only to the formal information conveyed by present clues (see point 4, below), it is not surprising that, in those problems, their preference for the positive-consistent responses was weakened.

2) In all of the studies, the formally correct responses were chosen significantly more often when they were consistent with the responses indicated by the present clues than when they were consistent with absent clues or were inconsistent with both present and absent clues (i.e., equiprobable problems).
3) In all of the studies, the mean confidence toward responses to the 18 problems, across participants, was positively correlated with the number of presence-consistent choices that were selected for those problems. It did not correlate significantly with the number of absence-consistent or equiprobable choices.
4) In all studies, the formal amount of information (as measured by ΔI) conveyed by the subset of present clues in each problem correlated positively with the number of presence-consistent choices on that problem, across participants. The formal amount of information conveyed by the subset of absent clues did not correlate significantly with the number of absence-consistent choices. These intriguing findings suggest that, although humans are probably sensitive to some extent to formal amounts of information (e.g., Cherubini et al., 2009; Oaksford & Chater, 1994), this is mostly the case when they evaluate the meaning of occurrences. Apparently, people can sometimes perceive that the absence of some features lends support to a hypothesis (actually, absence-consistent correct choices were not frequent, but they were not totally absent); however, in those instances they are, on average, at a loss for establishing *how much* support those absent clues lend to the hypothesis.

These converging pieces of evidence across the three different experiments were also confirmed in their conjoint analysis. Hence, they are mostly independent of the presentation format of the probabilities of the clues under the two alternative hypotheses, which was manipulated across the three experiments. They are also mostly independent of the ratio of present-to-absent features presented in each problem, which was manipulated within each experiment.

Ancillary findings: Possible moderators of the feature-positive effect

Rarity of the absent clues

Rarity effects concern the apportionment of increased attention to rare events in contrast to common ones (e.g., Feeney, Evans & Clibbens, 2000; Feeney, Evans & Venn, 2008; Green & Over, 2000; McKenzie & Mikkelsen, 2000, 2007; Oaksford & Chater, 1994; 2003; in legal contexts, for example, see Loftus, 1976; Wells & Lindsay, 1980). We included in our initial predictions a hypothesis that was based on rarity effects, conjecturing that participants would possibly pay more heed to absent clues when they were rare in comparison to present ones. The prediction followed from Newman et al. (1980)'s evolutionary-based argument that feature-positive effects originate from the fact that, in nature, occurrences are less common than non-occurrences and thus are, in a very general sense, more informative. Following from that argument, in specific contexts in which absent clues occur less than present clues, an opposite trend to pay heed to absent clues could arise. Accordingly, we devised different versions of each problem, varying the ratio of present-to-absent clues along three levels (2:2; 3:2; 2:3). Results were inconclusive with respect to the original prediction. A slight weakening of the feature positive effects occurred in Experiment 1 in the 3:2 problems, as shown by the type of problems × ratio of present-to-absent two-way interaction for the frequency of correct responses observed in that experiment. However, the interaction, although it was still significant, followed a distinctively different pattern in Experiment 2 and was not significant in Experiment 3 (thus giving rise to the three-way interaction observed in the cross-experimental analyses of correct responses). The cross-experimental analyses of the presence-consistent responses showed a decrease of presence-consistent choices occurring in the 3:2 equiprobable problems only, that is, in those problems in which the present clues were least informative. This set of different findings suggests that the rarity of absent clues might, in some circumstances, draw attention to them, but this effect is not systematic, and it apparently interacts with the presentation format of probabilistic information as well as with the formal amount of information conveyed by the stimuli in ways that are in need of further specification.

Presentation format of the probabilistic information

In most past experiments on hypothesis testing and evaluation that used explicit probabilistic information, only the probabilities of the occurrences of different features

were communicated to participants (e.g., Cherubini et al., 2010; McKenzie, 2006; Skov & Sherman, 1986; Slowiaczek et al., 1992). We conjectured that this format might inflate feature-positive effects, because the probabilities of non-occurrences have to be inferred by complementation. Accordingly, we systematically changed the way probabilities were communicated to the participants across the three experiments: in Experiment 1, we communicated the probabilities of occurrences; in Experiment 2, we communicated the probabilities of occurrences and non-occurrences; in Experiment 3, we communicated exclusively the probabilities of non-occurrences. Contrary to the initial conjecture, the probabilistic format did not have appreciable effects on choices, as shown by the cross-experimental analyses. However, it had some effect on confidence ratings. In Experiment 1, participants trusted presence-consistent responses more than all other responses, whereas, in Experiments 2 and 3, confidence toward presence-consistent and absence-consistent responses was similar (however, the correlation of confidence with the number of presence-consistent choices in each problem, across participants, was significant in all experiments). Apparently, communicating explicitly the probabilities of non-occurrences gives a hint to the participants that those probabilities should be considered and thus decreases their trust toward responses based only or mostly on the present clues. However, it does not actually help participants to correctly consider absent clues in their eventual decision.

The only other appreciable effect of the different probabilistic formats is its interaction with the rarity of absent features, as discussed in the previous paragraph.

Conclusion

The present scrutiny shows that, in the evaluation stage of hypothesis development, the occurrence of clues is systematically overrated with respect to the non-occurrence of other clues. The tendency to neglect the significance of the dog that failed to bark, as noted by Arthur Conan Doyle and mentioned by Ross (1978) and Fischhoff and Beyth-Marom (1983), far from being supported only by anecdotes, can be robustly observed in abstract laboratory tasks with fully explicit probabilistic information. The tendency is not appreciably or systematically weakened in those contexts in which non-occurrences are rare in comparison to occurrences, neither is it by the overt display of the probabilities of non-occurrences. Furthermore, this feature-positive effect influences confidence towards judgments: On average, participants trusted judgments based on occurrences more than

those based on non-occurrences (although this effect was cancelled by the overt presentation of the probabilities of non-occurrence). Finally, participants showed a remarkable sensitivity to the formal amount of information conveyed by the occurrence of stimuli (as shown by a positive correlation between the ΔI of and the number of presence-consistent choices), but they were insensitive to the formal amount of information conveyed by non-occurrences.

The feature-positive effect in hypothesis evaluation might have important consequences for confirmation biases (e.g., Klayman, 1995; McKenzie, 2004, 2006). The most common information-gathering testing strategy is positive testing, consisting of the search for clues whose occurrence is consistent with the hypothesis under examination. That is, when a hypothesis is tested positively, occurrences confirm it, whereas non-occurrences confute it. This information-gathering strategy, if coupled with the tendency to overrate occurrences in comparison to non-occurrences, would give rise to a systematic tendency to improperly confirm the tested hypothesis (e.g., Klayman, 1995; McKenzie, 2004, 2006). This type of confirmation bias might have important, detrimental side effects in contexts in which the rigorous testing of hypotheses is of critical importance, such as in scientific research, forensic practice, medical diagnosis (e.g., Christensen-Szalansky and Bushyhead's 1981; Scandellari, 2005) and health behaviors (e.g., Rassin, Muris, Franken, & van Straten, 2008). It might also have important consequences in the social domain, where feature-positive effects have been proven to occur (e.g., Fazio, Sherman, & Herr, 1982) and other types of confirmatory tendencies toward stereotypes are already known (e.g., Fiedler & Walther, 2004). However, because the present experiments used abstract problems only, estimating the impact of the feature-positive effect on the evaluation of hypotheses in practical domains will require further investigation.

Finally, it has yet to be clarified whether and under what circumstances the feature-positive effect might diminish or even reverse to a feature-negative effect (FNE, Fiedler, Eckert, & Poysiak, 1989) when evaluating competing hypotheses.

Conclusions

Collecting and interpreting the information retrieved from the environment or recalled from memory are activities which play an essential, maybe underrated, role in our lives, especially in this world which is “a highly uncertain place” (Chater & Oaksford, 2008). Specific situations, such as those of the physicians who make a diagnosis or judges who decide whether or not to sentence a defendant, or more common and mundane circumstances, like getting acquainted, entail dealing with pieces of evidence with a certain degree of fallibility and which convey a certain amount of uncertainty. Often, we look for and evaluate new evidence without even noticing whether and to what extent we are guided by our prior expectancies (e.g., stereotypes) or biased by our motivations and desires, or influenced by hardwired psychological mechanisms. The result is that we are constantly at risk of losing part of the information which is actually available to us. Under some circumstances this loss of information is harmless, “errors” are costless or even adaptive in light of the contingent environment, while in other circumstances inappropriate information gathering and information processing lead to errors which bear severe consequences. Isolating the cognitive processes at the base of these thinking activities in order to understand how they work and eventually to prevent maladaptive and pernicious consequences, as well to promote the adaptive and beneficial outcomes, is the final aim of the experiments that psychologists conduct in their laboratories. Reproducing through simple tasks the complexity of our world allows us to observe people’s behavior and systematically determine what influences it, eventually fostering the building of models of the mental activities which underlie it. A further step requires to get out from the laboratories and pit the model against the reality by understanding how and why people behave in a certain fashion in real-world situations.

The experiments we have illustrated in the previous chapters fall within this framework and aimed at further deepening our knowledge about people’s information-search and information-processing behavior when they are engaged in abstract tasks, which might be thought of as miniatures of real-world situations. Much has yet to be done in terms of building adequate models of testers’ behaviors as well as of assessing them in

real-world contexts. However, this fifty years of basic research in this field, if anything, allowed to rise the debate about human “rationality”/”irrationality”, and they paved the way to the implementation of increasingly improved procedures to foster efficient judgment and decision making in crucial domains, such as the medical or the judicial ones, where inferential errors are costly.

Where we were, the state of affairs and future directions

We have focused our empirical investigation on information gathering and information evaluation, two of the major steps of hypothesis development (e.g., McKenzie, 2004). As to the information-gathering process in abstract probabilistic tasks of hypothesis testing, there was a general consensus in previous literature around the notion that three main factors seem to drive people’s question-asking preferences: diagnosticity, positivity, and “extremity” (e.g., McKenzie, 2004; Slowiaczek et al., 1992). The seminal studies by Joshua Klayman, Steven J. Sherman, Richard B. Skov and co-workers at the end of the 80s and the beginning of the 90s (Skov & Sherman, 1986; Slowiaczek et al., 1992) seemed enough to give one pause for thought. The results of our experiments challenge some of their conclusions. Indeed, the upshot of our investigation on people’s strategies in abstract information gathering (see Chapter 2) is that people are indeed influenced by both the positivity and the diagnosticity of questions, but neither by their symmetry or they asymmetry (i.e., “extremity”, in its extended definition). Furthermore, the lack of a significant correlation between p (E | ¬H) and the mean rank assigned to each of the 32 questions to evaluate (Experiment 4, Chapter 2) suggests that participants were not trying to maximize the probability of occurrence of a feature *per se*, that is, p (E), rather they proved to be affected by the magnitude of the probability of occurrence of the feature *under the working hypothesis*, that is, p (E | H). This finding runs counter to the interpretation given by Skov & Sherman (1986) for the alleged preference for extreme (i.e., asymmetrically disconfirming) tests in terms of confirmation bias, defined as a tendency to maximize the probability of occurrence of the confirming outcome. Both the lack of preference for “extreme” tests and the lack of correlation of participants’ responses with p (E) in our Experiment 4 (Chapter 2) indicate that one of the three testing-evaluation combinations, namely extremity coupled with insensitivity to answer diagnosticity, described by Klayman (1995; see also McKenzie, 2004, 2006) as leading to confirming behaviors might not hold on firm ground. Of course, our finding needs to be

replicated, especially so in more contextualized tasks. Indeed, some preliminary findings of a set of experiments that Simona Sacchi and colleagues are carrying out on social inferences point to the possibility that people might deem more useful the frequency of the confirming evidence rather than its diagnostic value, in keeping with Skov & Sherman's (1986) interpretation. Yet, our results, if anything, emphasize the need of further empirical investigations on this topic. Indeed, before working out in depth the issue of *why* people behave in a certain fashion, it seems that there is still to do in the way of clarifying *what* is people's behavior in hypothesis-testing tasks.

The same reasoning might apply to the investigation of how people use incoming evidence. Scant research has been devoted to this topic, and it has shown that people tend to perceive the "yes" and "no" answers to a dichotomous question as more equally informative than they actually are. Hence, people are relative insensitive to differentially diagnostic answers, at least when the task has an abstract content (McKenzie, 2006; Skov & Sherman, 1986; Slowiaczek et al., 1992). The explanation given to this phenomenon rested upon a confusion of a procedure which turns out to be useful for assessing question diagnosticity, namely the feature-difference heuristic (Nelson, 2005, footnote 2; Nelson et al., 2010; Slowiaczek et al., 1992), as a good strategy to assess answer diagnosticity as well (Slowiaczek et al., 1992). The experiments we have presented in Chapter 3 adds to the literature by showing that the insufficient sensitivity to differentially diagnostic answers might be better conceived as a more general failure in information use. Specifically, it turned out that not only people have difficulty to perceive differences in informativeness of different answers to the same question, but they also tend to weigh as differentially diagnostic "yes" and "no" answers which are equally diagnostic (i.e., answers to a symmetric query), thus exhibiting *oversensitivity* to answer diagnosticity. This finding runs counter both to the interpretation of the phenomenon given by Slowiaczek et al. (1992), because people should be calibrated in that the difference in likelihoods are the same for "yes" and "no" answers when the tests are symmetric, and to their claim that the evaluation of the answers following symmetric questions cannot lead to inferential errors (Slowiaczek et al., 1992). On the contrary, if oversensitivity to equally diagnostic answers following a symmetric test manifests in terms of weighting more the "yes" answer than the "no" answer (as in our Experiment 1, Chapter 3), then the combination of positive-symmetric testing and oversensitivity would lead to an unwarranted overreliance in the working hypothesis, that is, to confirmation bias.

By contrast, insensitivity to answer diagnosticity might have a falsifying effect when the answers to evaluate follow an asymmetrically confirming query. Indeed, it turned out that participants to our experiments tended to perceive the “yes” answer to such questions as more similar in informativeness to the “no” answer than normatively expected. Hence, when one use an asymmetrically confirming test as a way to maximize the diagnostic value of the confirming outcome, the insensitivity in the evaluation stage leads to perceive the confirming outcome as less diagnostic than it actually is, thus reducing the extent of the confirmation tendency.

In terms of explaining people’s failure to perceive answer informativeness, beyond providing evidence against the use of the feature-difference heuristic, we found that people might over-rely on the given information, in particular on the prior probabilities of the hypotheses and, mimicking what we found in the testing phase (see Experiment 4, Chapter 2), the probability of a feature under the working hypothesis. Furthermore, adding the probabilities of non-occurrence of the features under the two hypotheses improved participants’ sensitivity to answer diagnosticity, suggesting that participants correctly focused on the given pieces of information, but then somehow failed to integrate them in a Bayesian fashion. Future studies should further address the issue of *why* people fail to correctly perceive the evidential strength of different answers to the same question by investigating whether the difficulty actually rests on integrating available information. However, a preliminary step should be to further elucidate the extent and the circumstances under which people fail to use information.

In line with the latter intent, the three experiments described in Chapter 4 specifically investigated a tendency that emerged both in our experiments and in previous studies on insensitivity to answer diagnosticity (e.g., Slowiaczek et al., 1992), namely the tendency to give more weight to occurrences than to non-occurrences, which might contribute to the understanding of why people are insensitive/oversensitive to differentially/equally diagnostic answers. What we found strongly support the existence of a feature-positive effect in hypothesis evaluation, wherein people apportion more attention to the presence of features than to their absence. Indeed, participants mostly preferred the hypothesis favored by the subset of present clues, irrespective of its actual correctness. However, they also exhibited sensitivity to the formal correct responses, even though only with respect to present clues, as shown by the correlations between the information gain (in

bits) conveyed by the present clues and the frequency of choices in line with an exclusive consideration of the present data.

This finding is interesting because it reflects a more general tendency which can be retrieved in all of the experiments we have presented in this contribution. Although participants were influenced by formally irrelevant properties, such as the probability of occurrence of a feature under the working hypothesis (both in the *testing* phase and in the *evaluation* phase of hypothesis development), or the presence of a feature (as opposed to its absence), they were also sensitive to formally relevant properties of the stimuli, such as the diagnosticity of questions and, as we have just reported, the formal amount of information conveyed by present data. This points to a consideration, which of course deserves further empirical investigations, about human ability to deal with information, either when gathering it or when interpreting it: It appears as thought people's behavior is not to be conceived as "irrational" or "rational" *tout court*. What emerges from our experiments is that people resort to heuristic processes (e.g., the focus on the probabilities under the focal hypothesis) or are driven by compelling psychological tendencies (e.g., overlooking of the absence of features), but occasionally perceive and use some formal properties of the task at hand (e.g., the priors or the information gain).

The latter point is related to an important aspect which future studies should take into account, that is, the choice of the normative model with which people's behavior should be compared. As noted by Jonathan Nelson (2005), a normative criterion might fail to capture participants' behavior (as a case in point, probability gain in the experiments by Baron et al., 1988), while others do not. In other words, a response that could be considered as an "error" when holding a certain normative criterion, might turn out to be perfectly "rational" when considering another normative model. Experiments devised to discriminate the descriptive power of competing normative models are thus in need both to describe and to explain participants' responses avoiding to rush to judgments about their rationality.

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
deck A	77	80	20	8
deck B	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 deck A *(deck B, 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 deck A *(deck B, 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 = .25 for “no”) and “Is there a pentagon?” (LR = .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.*

Problem	*Deck*	*p(triangle)*	*p(circle)*	*p(square)*	*p(pentagon)*
1	*A*	.8	.9	.08	.2
	B	.2	.08	.9	.8
2	*A*	.8	.46	.05	.2
	B	.2	.05	.46	.8
3	*A*	.8	.9	.1	.3
	B	.3	.1	.9	.8
4	*A*	.95	.8	.2	.54
	B	.54	.2	.8	.95
5	*A*	.8	.48	.12	.2
	B	.2	.12	.48	.8
6	*A*	.8	.88	.52	.2
	B	.2	.52	.88	.8
7	*A*	.8	.92	.23	.2
	B	.2	.23	.92	.8
8	*A*	.77	.8	.2	.08
	B	.08	.2	.8	.77

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

Problem	*Question*	*Response*	*p(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				
	Pentagon?	*Yes*	.5	.28	6.02	.28	6.02	Symmetric, negative	Symmetric, positive
		No	.5	.28	6.02				
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				
	Pentagon?	*Yes*	.5	.28	6.02	.28	6.02	Symmetric, negative	Symmetric, positive
		No	.5	.28	6.02				
3	*Triangle?*	*Yes*	.55	.15	4.26	.19	4.79	Disconfirming, positive	Confirming, negative
		No	.45	.24	5.44				
	Circle?	*Yes*	.5	.53	9.54	.53	9.54	Symmetric, positive	Symmetric, negative
		No	.5	.53	9.54				
	Square?	*Yes*	.5	.53	9.54	.53	9.54	Symmetric, negative	Symmetric, positive
		No	.5	.53	9.54				
	Pentagon?	*Yes*	.55	.16	4.26	.2	4.79	Confirming, negative	Disconfirming, positive
		No	.45	.24	5.44				
4	*Triangle?*	*Yes*	.75	.06	2.45	.18	4.29	Disconfirming, positive	Confirming, negative
		No	.26	.53	9.64				
	Circle?	*Yes*	.5	.28	6.02	.28	6.02	Symmetric, positive	Symmetric, negative
		No	.5	.28	6.02				
	Square?	*Yes*	.5	.28	6.02	.28	6.02	Symmetric, negative	Symmetric, positive
		No	.5	.28	6.02				
	Pentagon?	*Yes*	.75	.06	2.45	.18	4.29	Confirming, negative	Disconfirming, positive
		No	.26	.53	9.64				
5	*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	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				
	Pentagon?	*Yes*	.5	.28	6.02	.28	6.02	Symmetric, negative	Symmetric, positive
		No	.5	.28	6.02				
6	*Triangle?*	*Yes*	.5	.28	6.02	.28	6.02	Symmetric, positive	Symmetric, negative
		No	.5	.28	6.02				
	Circle?	*Yes*	.7	.05	2.28	.12	3.41	Disconfirming, positive	Confirming, negative
		No	.3	.28	6.02				
	Square?	*Yes*	.7	.05	2.28	.12	3.41	Confirming, negative	Disconfirming, positive
		No	.3	.28	6.02				
	Pentagon?	*Yes*	.5	.28	6.02	.28	6.02	Symmetric, negative	Symmetric, positive
		No	.5	.28	6.02				
7	*Triangle?*	*Yes*	.5	.28	6.02	.28	6.02	Symmetric, positive	Symmetric, negative
		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				
	Pentagon?	*Yes*	.5	.28	6.02	.28	6.02	Symmetric, negative	Symmetric, positive
		No	.5	.28	6.02				
8	*Triangle?*	*Yes*	.43	.55	9.83	.39	7.64	Confirming, positive	Disconfirming, negative
		No	.58	.28	6.02				
	Circle?	*Yes*	.5	.28	6.02	.28	6.02	Symmetric, positive	Symmetric, negative
		No	.5	.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(resp) = probability of receiving that response

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

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

Expected I.G.= p(yes)*I.G.(yes)+p(no)*I.G.(no). A measure of the diagnosticity of a question

Mean W.E. = p(yes)*W.E.(yes)+p(no)*W.E.(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.*

Problem	*Deck*	*p(triangle)*	*p(circle)*	*p(square)*	*p(pentagon)*
1	*A*	.82	.92	.06	.18
	B	.18	.06	.92	.82
2	*A*	.81	.47	.04	.19
	B	.19	.04	.47	.81
3	*A*	.81	.88	.12	.39
	B	.39	.12	.88	.81
4	*A*	.96	.86	.14	.37
	B	.37	.14	.86	.96
5	*A*	.82	.41	.09	.18
	B	.18	.09	.41	.82
6	*A*	.84	.92	.58	.16
	B	.16	.58	.92	.84
7	*A*	.79	.94	.25	.21
	B	.21	.25	.94	.79
8	*A*	.75	.79	.21	.06
	B	.06	.21	.79	.75

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

Problem	*Question*	*Response*	*p(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				
	Pentagon?	*Yes*	.5	.32	6.59	.32	6.59	Symmetric, negative	Symmetric, positive
		No	.5	.32	6.59				
2	*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				
	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				
3	*Triangle?*	*Yes*	.6	.09	3.17	.14	3.93	Disconfirming, positive	Confirming, negative
		No	.4	.20	5.07				
	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				
4	*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				
	Pentagon?	*Yes*	.665	.14	4.14	.32	6.76	Confirming, negative	Disconfirming, positive
		No	.335	.67	11.97				
5	*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				
	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				
6	*Triangle?*	*Yes*	.5	.37	7.2	.37	7.2	Symmetric, positive	Symmetric, negative
		No	.5	.37	7.2				
	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				
7	*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				
	Pentagon?	*Yes*	.5	.26	5.75	.26	5.75	Symmetric, negative	Symmetric, positive
		No	.5	.26	5.75				
8	*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				
	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				

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

Problem	Deck	P(triangle)	P(circle)	P(square)	P(pentagon)	Diagnosticity of questions
1	A	.11	**.98**	.89	**.02**	Low (mean WE = .87, I.G. between .026 and .008)
	B	**.02**	.89	**.98**	.11	
2	A	**.68**	.11	.58	**.02**	
	B	.58	**.02**	**.68**	.11	
3	A	**.98**	**.68**	.89	.58	
	B	.89	.58	**.98**	**.68**	
4	A	**.68**	.49	.58	**.39**	
	B	.58	**.39**	**.68**	.49	
5	A	**.98**	.38	**.02**	.62	High (mean W.E. = 4.15, I.G. between .15 and .17)
	B	.62	**.02**	.38	**.98**	
6	A	.38	.62	**.02**	**.18**	
	B	**.02**	**.18**	.38	.62	
7	A	.62	**.98**	**.18**	.62	
	B	**.18**	.62	.62	**.98**	
8	A	.62	**.88**	.45	**.18**	
	B	**.18**	.45	**.88**	.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: $p > .9$

High: $.5 < p < .9$

Low: $.5 > p > .1$

Very low: $p < .1$

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(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.

Appendix H. *Sample stimulus from Experiments 1 and 2.*

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

Study 1 version:

	Have gills
Gloms	65%
Fizos	35%

Study 2 version:

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

The creature is asked: "Do you have gills?".

It answers: "Yes, I do".

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

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

Appendix I. *The distribution of probabilities of each letter in each deck for the 18 problems used in the three experiments (in bold the probabilities of the present clues).*

Problem	*Deck*	*p(B)*	*p(C)*	*p(D)*	*p(F)*	*p(G)*
1	*1*	**.43**	**.8**	.89	.93	
	2	**.01**	**.08**	.1	.1	
2	*1*	**.03**	**.29**	.35	.65	.25
	2	**.3**	**.2**	.9	.35	.62
3	*1*	**.85**	**.8**	**.95**	.95	.96
	2	**.04**	**.44**	**.3**	.1	.1
4	*1*	**.35**	**.2**	.14	.4	
	2	**.1**	**.98**	.39	.4	
5	*1*	**.01**	**.11**	**.8**	.3	.2
	2	**.8**	**.75**	**.76**	.96	.9
6	*1*	**.9**	**.7**	.95	.96	
	2	**.02**	**.22**	.1	.1	
7	*1*	**.9**	**.7**	.9	.9	.9
	2	**.02**	**.22**	.2	.4	.2
8	*1*	**.5**	**.7**	**.3**	.5	.35
	2	**.09**	**.88**	**.97**	.4	.26
9	*1*	**.02**	**.16**	.1	.1	
	2	**.5**	**.7**	.94	.96	
10	*1*	**.8**	**.2**	.15	.45	.85
	2	**.07**	**.68**	.4	.35	.4
11	*1*	**.02**	**.16**	**.5**	.1	.05
	2	**.7**	**.6**	**.45**	.95	.95
12	*1*	**.09**	**.88**	.85	.16	
	2	**.1**	**.5**	.2	.75	
13	*1*	**.02**	**.16**	.2	.12	.1
	2	**.5**	**.7**	.9	.9	.85
14	*1*	**.09**	**.88**	**.97**	.3	.76
	2	**.83**	**.22**	**.55**	.8	.35
15	*1*	**.01**	**.11**	.15	.1	.1
	2	**.85**	**.6**	.8	.8	.9
16	*1*	**.85**	**.65**	**.65**	.89	.94
	2	**.02**	**.16**	**.3**	.1	.05
17	*1*	**.01**	**.11**	.16	.15	
	2	**.75**	**.5**	.9	.95	
18	*1*	**.8**	**.5**	.6	.7	.96
	2	**.01**	**.08**	.3	.15	.1

Appendix J. Some properties of the 18 problems used in the three experiments.

Problem	*Correct*	*Suggested by present clues*	*Suggested by absent clues*	*I.G. present clues*	*I.G. absent clues*
1	*1*	*1*	*2*	.98	.92
2	*equiprobable*	*2*	*1*	.45	.45
3	*2*	*1*	*2*	.93	.98
4	*equiprobable*	*2*	*1*	.02	.02
5	*2*	*2*	*1*	.98	.94
6	*2*	*1*	*2*	.94	.98
7	*2*	*1*	*2*	.94	.97
8	*equiprobable*	*1*	*2*	.02	.02
9	*1*	*2*	*1*	.93	.97
10	*equiprobable*	*1*	*2*	.22	.22
11	*1*	*2*	*1*	.93	.97
12	*equiprobable*	*1*	*2*	.04	.04
13	*1*	*2*	*1*	.93	.98
14	*equiprobable*	*2*	*1*	.01	.01
15	*2*	*2*	*1*	.98	.95
16	*1*	*1*	*2*	.97	.94
17	*2*	*2*	*1*	.97	.94
18	*1*	*1*	*2*	.98	.93

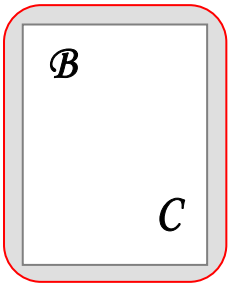
Appendix K. *Sample problems from the three experiments.*

Table K.1 *Sample problems from the Experiment 1.*

On the drawn card there are a B and a C, but not a D, a F and a G.

Put a mark within the box indicating the deck from which the card was most likely drawn.

	B	C	D	F	G
deck 1	3	29	35	65	25
deck 2	30	20	90	35	62



equiprobable	deck 1	deck 2

Mark your degree of confidence on the response you gave:

not confident 1 2 3 4 5 6 7 very confident

Table K.2 *Sample problems from the Experiment 2. The probabilistic information in Experiment 2 were displayed in the following way:*

	B		C		D		F		G	
	yes	no	yes	no	yes	no	yes	no	yes	no
deck 1	3	97	29	71	35	65	65	35	25	75
deck 2	30	70	20	80	90	10	35	65	62	38

Table K.3 *Sample problems from the Experiment 3. The instructions made that the numbers in the table indicated the number of cards—out of 100 in each deck—that did not display each letter:*

	B	C	D	F	G
deck 1	97	71	65	35	75
deck 2	70	80	10	65	38

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

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

Baron, J., Beattie, J., & Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning: II. Congruence, information, and certainty. *Organizational Behavior and Human Decision Processes,42*, 88-110. doi:10.1016/0749-5978(88)90021-0

Beyth-Marom, R. (1982). Perception of correlation reexamined. *Memory & Cognition, 10*, 511-519.

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

Bourne, L. E. Jr. & Guy, D. E. (1968). Learning conceptual rules. II: The role of positive and negative instances. *Journal of Experimental Psychology, 77*, 488-494. doi:10.1037/h0025952

Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (in press). Looking for Honesty: The Primary Role of Morality (vs. Sociability and Competence) in Information Gathering. *European Journal of Social Psychology*. doi: 10.1002/ejsp.744.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: John Wiley & Sons, Inc.

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

Chater, N., & Oaksford, M. (2008). The probabilistic mind: Prospects for a Bayesian cognitive science. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 3-31). Oxford, UK: Oxford University Press.

Chapman, L. J., & Chapman, J. P. (1959). Atmosphere effect re-examined. *Journal of Experimental Psychology, 58*, 220-226. doi:10.1037/h0041961

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17*, 391-416. doi:10.1016/0010-0285(85)90014-3

Cherubini, P. (2007). Fallacie nel ragionamento probatorio. In L. de Cataldo Neuburger (Ed.), *La prova scientifica nel processo penale* (pp. 249-302). Padova: CEDAM.

Cherubini, P., Castelvecchio, E., & Cherubini, A. M. (2005). Generation of hypotheses in Wason's 2-4-6 task: An information theory approach. *Quarterly Journal of Experimental Psychology, 58*, 309-332. doi:10.1080/02724980343000891

Cherubini, P., Rusconi, P., Russo, S., & Crippa, F. (submitted). Missing the dog that failed to bark in the nighttime: On the overestimation of occurrences over non-occurrences in hypothesis testing.

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

Cherubini, P., Russo, S., Rusconi, P., D'Addario, M., & Boccuti, I. (2009). Il ragionamento probabilistico nella diagnosi medica: sensibilità e insensibilità alle informazioni. In P. Giaretta, A. Moretto, G. F. Gensini, & M. Trabucchi (Eds.), *Filosofia della medicina: Metodo, modelli, cura ed errori* (pp. 541-564). Bologna: Il Mulino.

Christensen-Szalanski, J. J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 928-935. doi:10.1037//0096-1523.7.4.928

Cohen, L. J. (1977). *The probable and the provable*. Oxford: Clarendon Press.

Crupi, V., Tentori, K., & Gonzalez, M. (2007). On Bayesian theories of evidential support: Normative and descriptive considerations. *Philosophy of Science, 74*, 229-252.

Dardenne, B., & Leyens, J.-Ph. (1995). Confirmation bias as a social skill. *Personality and Social Psychology Bulletin, 21*, 1229-1239. doi:10.1177/01461672952111011

Devine, P. G., Hirt, E. R., & Gehrke, E. M. (1990). Diagnostic and confirmation strategies in trait hypothesis testing. *Journal of Personality and Social Psychology, 58*, 952-963. doi: 10.1037/0022-3514.58.6.952

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

Evans, J. St. B. T. (1972). Interpretation and matching bias in a reasoning task. *Quarterly Journal of Experimental Psychology, 24*, 193-199. doi: 10.1080/00335557243000067

Evans, J. St. B. T. (1998). Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking and Reasoning, 4*, 45-82.

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

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review, 83*, 37-64. doi:10.1037//0033-295X.83.1.37

Fazio, R. H., Sherman, S. J., & Herr P. M. (1982). The feature-positive effect in the self-perception process: Does not doing matter as much as doing? *Journal of Personality and Social Psychology, 42*, 404-411. doi:10.1037/0022-3514.42.3.404

Feeney, A., Evans, J. & Venn, S. (2008). Rarity, pseudodiagnosticity and Bayesian reasoning. *Thinking & Reasoning, 14*, 209-230. doi:10.1080/13546780801934549

Feeney, A., Evans, J. St. B. T., & Clibbens, J. (2000). Background beliefs and evidence interpretation. *Thinking and Reasoning*, 6, 97-124.Fiedler, K, Eckert, C., & Poysiak, C. (1989). Asymmetry in human discrimination learning: Feature positive effect or focus of hypothesis effect? *Acta Psychologica, 70*, 109-127. doi:10.1016/0001-6918(89)90015-2

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. doi:10.1037//0096-3445.129.3.399

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

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

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. doi:10.1006/jesp.2000.1441

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

Good, I.J. (1983). *Good thinking: the foundations of probability and its applications*. Minneapolis, MN: University of Minnesota Press.

Green, D. W., & Over, D. E. (2000). Decision theoretic effects in testing a causal conditional. *Current Psychology of Cognition*, 19, 51–68

Hammond, K. R. (2007). *Beyond rationality: The search for wisdom in a troubled time*. Oxford: Oxford University Press.

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

Hearst, E., & Wolff, W. T. (1989). Addition versus deletion as a signal. *Animal Learning & Behavior, 17*, 120-133.

Henrion, M., & Fischhoff, B. (1986). Assessing uncertainty in physical constants. *American Journal of Physics, 54*, 791-798. doi:10.1119/1.14447

Hovland, C. I., & Weiss, W. (1953). Transmission of information concerning concepts through positive and negative instances. *Journal of Experimental Psychology, 45*, 175-182. doi:10.1037/h0062351

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434-446. doi:10.1016/j.jml.2007.11.007

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

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

Jones, M., & Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision, 50*, 59-99.

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

Kao, S-F., & Wasserman, E. A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1363-1386. doi:10.1037//0278-7393.19.6.1363

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.

Kirby, K. N. (1994). Probabilities and utilities of fictional outcomes in Wason's four-card selection task. *Cognition 51,* 1-28. doi:10.1016/0010-0277(94)90007-8

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

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*, 211-228. doi: 10.1037/0033-295X.94.2.211

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. doi:10.1037//0278-7393.15.4.596

Koehler, D. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes, 56*, 28-55. doi:10.1006/obhd.1993.1044

Loftus, E. F. (1976). Unconscious transference in eyewitness identification. *Law and Psychology Review, 2*, 93-98.

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. doi:10.1037//0022-3514.37.11.2098

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. doi:10.1037//0096-3445.127.3.269

McDonald, J. (1992). Is strong inference superior to simple inference? *Syntheses,* 92, 261-282.

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

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

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

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

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

Nahinsky, I. D., & Slaymaker, F. L. (1970). Use of negative instances in conjunctive concept identification. *Journal of Experimental Psychology, 84*, 64-68. doi:10.1037/h0028951

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

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

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

Newell, B. R. (2009). A science of decision making: The legacy of Ward Edwards. J.W. & D.J. Weiss (Eds.). Oxford University Press, New York. *Journal of Economic Psychology, 30*, 694-695. doi:10.1016/j.joep.2009.05.001

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

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

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2,* 175-220. doi: 10.1037/1089-2680.2.2.175

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. doi:10.1037//0033-295X.101.4.608

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, UK: 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.] (2001). *Hypothesis-testing behaviour*. Hove, U.K.: Psychology Press.

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

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.

Rakow, T. (2010). Risk, uncertainty and prophet: The psychological insights of Frank H. Knight. *Judgment and Decision Making, 5*, 458-466.

Rassin, E., Muris, P, Franken, I., & van Straten, M. (2008). The feature-positive effect and hypochondriacal concerns. *Behaviour Research and Therapy, 46*, 263–269. doi:10.1016/j.brat.2007.11.003

Reverberi, C., Rusconi, P., Paulesu, E., & Cherubini, P. (2009). Response demands and the recruitment of heuristic strategies in syllogistic reasoning. *Quarterly Journal of Experimental Psychology, 62*, 513-530. doi: 10.1080/17470210801995010.

Riva, P., Rusconi, P., Montali, L., & Cherubini, P. (in press). The influence of anchoring on pain judgment. *Journal of Pain and Symptom Management*.

Ross, L. (1978). The intuitive psychologist and his shortcomings: distortions in the attribution process. In L. Berkowitz (Ed.), *Cognitive theories in social psychology: papers from advances in experimental social psychology*. New York: Academic Press.

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. doi:10.1080/02724980042000101

Rusconi, P., & McKenzie, C. R. M. (in preparation). Testing different accounts of insensitivity and oversensitivity to differentially informative answers in abstract hypothesis testing.

Rusconi, P., Riva, P., Cherubini, P., & Montali, L. (2010). Taking into account the observers' uncertainty: a graduated approach to the credibility of the patient's pain evaluation. *Journal of Behavioral Medicine, 33*, 60-71. doi: 10.1007/s10865-009-9232-5.

Sacchi, S., Rusconi, P., Russo, S., Bettiga, R., & Cherubini, P. (accepted with minor revision). New knowledge for old credences: Asymmetric information search about in-group and out-group members. *British Journal of Social Psychology*.

Scandellari, C. (2005). *La diagnosi clinica: Principi metodologici del procedimento decisionale*. Milano: Masson.

Sells, S. B. (1936). The atmosphere effect: An experimental study of reasoning. *Archives of Psychology, 29,* 3-72.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*, 379-423, 623-656.

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

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.

Smedslund, J. (1963). The concept of correlation in adults. *Scandinavian Journal of Psychology, 4*, 165-173. doi:10.1111/j.1467-9450.1963.tb01324.x

Snyder, M., & Swann, W. B. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology, 36*, 1202-1212. doi:10.1037/0022-3514.36.11.1202

Spellman, B.A., Lopez, A., & Smith, E.E. (1999). Hypothesis testing: Strategy selection for generalising versus limiting hypotheses. *Thinking and Reasoning, 5*, 67-91. doi:10.1080/135467899394084

Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition, 57*, 31-95. doi:10.1016/0010-0277(95)00666-M

Tabak, J. (2004). *Probability and statistics: The science of uncertainty.* New York: Facts On File, Inc.

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

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

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

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

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.

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

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

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

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *The Quarterly Journal of Experimental Psychology, 12*, 129-140. doi:10.1080/17470216008416717

Wason, P. C. (1961). Response to affirmative and negative binary statements. *British Journal of Psychology, 52*, 133-142.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology I*. Harmondsworth, UK: Penguin.

Wason, P.C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*, 273-281.

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

Wetherick, N. E. (1962). Eliminative and enumerative behaviour in a conceptual task. *Quarterly Journal of Experimental Psychology, 14,* 246-249. doi: 10.1080/17470216208416542

Woodworth, R. S., & Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology, 18,* 451-460. doi:10.1037/h0060520

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. doi: 10.1037/0022-3514.68.1.52