Integration of base rates and new information in an abstract hypothesis-testing task

Patrice Rusconi[a], Marco Marelli[a], Selena Russo[b], Marco D'Addario[a], Paolo Cherubini[a]

[a] University of Milano-Bicocca

[b] University of Trento

Author Note

Patrice Rusconi, Department of Psychology, University of Milano-Bicocca; Marco Marelli, Department of Psychology, University of Milano-Bicocca; Selena Russo, Department of Cognitive Sciences and Education, University of Trento; Marco D'Addario, Department of Psychology, University of Milano-Bicocca; Paolo Cherubini, Department of Psychology, University of Milano-Bicocca.

Selena Russo is now in Sydney, Australia.

Correspondence concerning this article should be addressed to Patrice Rusconi, Department of Psychology, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milano, Italy. Phone: ++39 02 6448 3736. Fax: ++39 02 6448 3706. E-mail: patrice.rusconi1@unimib.it

Abstract

In two studies, we investigated how people use base rates and the presence vs. the absence of new information to judge which of two hypotheses is more likely. Participants were given problems based on two decks of cards printed with 0 to 4 letters. A table showed the relative frequencies of the letters on the cards within each deck. Participants were told the letters that were printed on or absent from a card the experimenter had drawn. Base rates were conveyed by telling participants that the experimenter had chosen the deck by drawing from an urn containing, in different proportions, tickets marked either "deck 1" or "deck 2". The task was to judge from which of the two decks the card was most likely drawn. Prior probabilities and the evidential strength of the subset of present clues (computed as "weight of evidence") were the only significant predictors of participants' dichotomous (both studies) and continuous (Study 2) judgments. The evidential strength of all clues was not a significant predictor of participants' judgments in either study, and no significant interactions emerged. We discuss the results as evidence for additive integration of base rates and the new *present* information in hypothesis testing.

Keywords: hypothesis testing; the feature-positive effect; additivity.

Integration of base rates and new information in an abstract hypothesis-testing task

Imagine that a woman arrives at an emergency room with widespread abdominal pain and fever. Based on this limited information, at least two alternative diagnoses, toxic infection and appendicitis may appear plausible to a physician. The fever suggests that the patient is suffering from an inflammatory condition whereas the abdominal pain indicates infection in a delimited area. The physician might consider a diagnosis of toxic infection slightly more probable than one of appendicitis based on a priori considerations, such as the incidence of the two diseases and the patients she/he usually sees. Although it is essential to consider both the disease incidence and the occurrences of symptoms during the diagnostic process, an efficient diagnosis also relies upon an appropriate evaluation of the absence of some specific medical signs. In the aforementioned scenario, for example, the physician should also take the facts that the patient does not show important symptoms such as vomiting and diarrhea into consideration. According to the toxic infection hypothesis, these symptoms should be present whereas they are often absent under the alternate hypothesis of appendicitis. Therefore, an accurate revision of the a priori considerations in light of a comprehensive evaluation of the entire body of evidence that is provided by both the present and absent symptoms would lead the clinician to be more confident in a diagnosis of appendicitis than in a diagnosis of toxic infection.

This scenario illustrates the way in which a Bayesian-like revision process whereby the base-rate information is accurately integrated with the "*indicant* or *diagnostic* information" (Bar-Hillel, 1980) should be carried out. The aim of the present study was to extend the results from previous literature regarding the ways in which people interpret both the presence and the absence of features and how they combine this information with the prior probabilities of the outcomes in abstract problems of hypothesis testing.

Since Kahneman and Tversky's (1973) seminal study that reported that people have a tendency to base judgments more on similarity (or representativeness) than on base rates, there have been a number of studies that have dealt with the way in which people use base-rate information

when making judgments. The predictions that the participants in this study made (e.g., estimating the probability that an individual is an engineer/lawyer) generally relied on the specific evidence that was available (e.g., personality sketches representative of either the stereotype of engineers or the stereotype of lawyers), and base-rate information (e.g., the composition of the set from which the sketches had been drawn) was rarely considered, even in cases in which the expected accuracy of any given prediction was low enough that the prior probabilities of the outcomes should have been weighted more heavily.

Several subsequent studies have confirmed that participants undervalue the prior probabilities of outcomes; this phenomenon is known as the base-rate fallacy (e.g., Casscells, Schoenberger, & Graboys, 1978; Fischhoff & Beyth-Marom, 1983; Lyon & Slovic, 1976; at the interpersonal level, see, e.g., Nisbett & Borgida, 1975). However, several studies have also elucidated the circumstances under which people consider and utilize base-rate information when making judgments (e.g., Ajzen, 1977; Bar-Hillel, 1980; Christensen-Szalanski & Bushyhead, 1981; Fischhoff, Slovic, & Lichtenstein, 1979; Ginosar & Trope, 1980; Ginossar & Trope, 1987; see Koehler, 1996 for a review)

In hypothesis-testing studies, it has been often assumed that the prior probabilities of various possible outcomes are equal (Fox & Rottenstreich, 2003; Nelson, 2005, footnote 12; Poletiek, 2001). Only a few studies of hypothesis testing[1] have considered the influence of unequal prior probabilities on the judgments that people make (Baron, Beattie, & Hershey, 1988; Nelson, 2005, footnote 12). To the best of our knowledge, the only study that specifically considered unequal prior probabilities in the context of hypothesis evaluation was done by Christensen-Szalanski and Bushyhead (1981). The participants in this study (physicians) demonstrated sensitivity to the prevalence of the disease that they were diagnosing (pneumonia); which corresponded to 3 cases *per* 100 patients in their clinical setting.

A less controversial phenomenon known as the feature-positive effect, which refers to the notion that people are more likely to pay attention to the presence of features than to their absence,

has been found in studies of several human and non-human perceptual and cognitive processes (e.g., Bourne & Guy, 1968; Hovland & Weiss, 1953; Nahinsky & Slaymaker, 1970; Neisser, 1963; Newman, Wolff, & Hearst, 1980; Treisman & Souther, 1985). Recent reports in the hypothesis-testing literature that focused on the hypothesis-evaluation stage of hypothesis development (see Klayman, 1995; McKenzie, 2004) have provided evidence that people tend to over-rely on occurrences and disregard non-occurrences in an abstract hypothesis evaluation task (Cherubini, Rusconi, Russo, & Crippa, in revision). However, evidence for this phenomenon was not found in Christensen-Szalanski and Bushyhead's (1981) study.

Bayesian background

Bayes' rule provides a normative criterion that is often used to weigh the impact that new information has on two or more competing hypotheses and to determine how the initial confidence that a decision-making entity has in each of these hypotheses (which is expressed via the prior probabilities of the competing hypotheses) should be adjusted in light of the new information. This criterion can be expressed in terms of odds by the following equation (e.g., Fischhoff & Beyth-Marom, 1983; Slovic & Lichtenstein, 1971):

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

where p () stands for "the probability of", ¬ is the logical symbol for negation, | is a logical symbol that means "given that", H is the hypothesis under consideration, $\neg H$ is the alternate hypothesis, and D is the set of all the pieces of evidence. Reading from the left of the formula, there are (1) the posterior odds, which is expressed as the ratio of the probability that the focal hypothesis is true given the acquired evidence to the probability that the alternate hypothesis is true given the same evidence; (2) the prior odds, which can be expressed as the ratio of the probability that the focal hypothesis is true to the probability that the alternate hypothesis is true prior to the receipt of the new evidence; and (3) the likelihood ratio (LR) of the probability of finding the evidence given the truth of the focal hypothesis to the probability of finding the same evidence given the truth of

the alternate hypothesis. In the case of absent information (or a "no" answer to a question), the LR is given by the following expression:

$$\frac{p(\neg D \mid H)}{p(\neg D \mid \neg H)}$$

where $\neg D$ indicates the missing data. Note that the conditional probability of a non-occurrence is complementary to the conditional probability of an occurrence as seen in the following expression:

$$\frac{p(\neg D \mid H)}{p(\neg D \mid \neg H)} = \frac{1 - p(D \mid H)}{1 - p(D \mid \neg H)}$$

In other words, from a formal (Bayesian) standpoint, the presence of a highly likely clue is tantamount to the absence of a highly unlikely clue (and vice versa), so any default preferences for knowledge about present over absent features are unwarranted.

Alan Turing (1912-1954) first used the log LR as a measure of the confirmatory/falsificatory strength of a datum (Good, 1979). In particular, he introduced the *ban* as the unit of measure of the *weight of evidence* (*WE*), which is determined according to the following expression (in *decibans*, that is, one-tenth of a *ban*):

$$WE = 10 \times \log_{10}\left[\frac{p(D \mid H)}{p(D \mid \neg H)}\right]$$

In this form, the algebraic sign (+ or −) indicates the respective value (confirmatory or falsificatory) of the presence of the evidence relative to a focal hypothesis. A value of 0 means that the evidence is uninformative.

Several other (Bayesian) measures of the confirmatory/falsificatory strength of a datum have been described in the literature (e.g., Crupi, Tentori, & Gonzalez, 2007; Nelson, 2005, 2008; Nelson, McKenzie, Cottrell, & Sejnowski, 2010). Unlike the *WE*, there are other metrics that take both the prior probabilities and the posterior probabilities into account. For example, the information gain (IG) measure, which is measured in *bits* and which is derived directly from Shannon's (1948) definition of entropy (or uncertainty), is expressed as:

$$E_n(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

where *X* is a discrete random variable that can assume x_i possible values, each of which has probability $p(x_i)$. The IG is the difference between the entropy prior to receiving the new information (*D*) and the entropy after receiving the new information, that is:

$$IG = \sum_{i=1}^{n} p(x_i)\log_2 p(x_i) - \sum_{i=1}^{n} p(x_i \mid D)\log_2 p(x_i \mid D)$$

Irrespective of the specific model that is chosen to measure the utility value of a datum, an efficient (Bayesian) revision of initial beliefs rests on the accurate, multiplicative integration of the information that is conveyed by the prior probabilities and the information transmitted by the likelihood ratios of present features and those of absent features. A non-normative integration of these two sources of information reflects different possible underlying difficulties, which were illustrated by Fischhoff and Beyth-Marom (1983): (a) "*misperception,*" which is defined as a failure to assess the actual diagnosticity of the new evidence (b) "*misaggregation,*" which is defined as a failure to integrate the information from each source that occurs because either a non-normative rule of composition (e.g., averaging instead of multiplying) has been applied or the multiplicative rule from Bayes' theorem has been correctly selected but erroneously applied.

In the following two studies, we aimed to clarify whether and how people integrate information in a one-shot laboratory task in which participants were asked to evaluate two hypotheses when the probability distributions of both present and absent evidence were given.

Overview of the studies

We devised two studies to examine the ways in which people interpret incoming evidence, which was either present or absent, and integrate it with relevant base-rate information to make a decision about two mutually exclusive and jointly exhaustive hypotheses in a one-shot task. Although some studies have investigated the degrees to which participants are sensitive to the evidential strength of incoming evidence in hypothesis evaluation (e.g., Cherubini et al., in revision;

McKenzie, 2006; Skov & Sherman, 1986; Slowiaczek, Klayman, Sherman, & Skov, 1992), to the best of our knowledge, the study by Christensen-Szalanski and Bushyhead (1981) is the only one that has both addressed the issue of how people revise their beliefs in one-shot tasks in light of either clues that were present or absent and considered the way in which participants used the unequal prior probabilities of the hypotheses. However, their study suffered from possible intercorrelations among the cues that were probably due to the realism of their task in terms of how it reflected the interrelationships between features in the real world. We attempted to extend their investigation to a pencil-and-paper hypothesis-testing task that used materials that were not explicitly representative of a real environment.

The materials and procedures of the two studies that are presented in this article were identical; these research methods were similar to those used by Cherubini et al. in another recent work (in revision). The only differences between the two present studies were in the dependent variables and, therefore, in a part of the instructions, for reasons we will discuss later.

Materials and procedure

In both studies, we presented the participants with 10 problems. All of the problems concerned two decks of cards, each of which was composed of 100 cards. Participants were told that each card within each deck showed between 0 and 4 letters on its face and that the letters had been chosen from the set {B, C, D, F}. In each problem, participants were given a table that showed the distributions of the letters, that is, participants were informed of the number of cards that had each letter printed on them within each deck. The instructions made it clear that the number of cards that depicted any given letter was independent of the number of cards that reported any other letter. Formally speaking, the letters on the cards represented the indicant information, that is, the likelihoods (see Appendix A for some of the formal properties of the 10 problems).

The participants were told that the experimenter had drawn a card from one of the two decks without disclosing the identity of the deck that he/she had chosen. However, they were given verbal and visual descriptions of the card that had been drawn; they were explicitly informed about which

letters were printed on the face of the card and which letters were absent (see Appendix B for a sample stimulus). In each problem, the drawn card always featured two letters, and the other two letters were absent. In all of the problems, the hypothesis that was supported by the subset of the present clues was the alternate of the hypothesis that was favored by the subset of the absent clues (see Appendix A). The participant was tasked with surmising the identity of the deck from which the card had most likely been drawn based on the presence vs. absence of the letters, the probability table that was associated with the problem, and the prior probabilities of the two hypotheses.

To provide the participants with relevant information about the prior probabilities of the two outcomes (i.e., deck 1 vs. deck 2), we told them that the experimenter had used a random process to choose the deck from which he/she then drew the card. Specifically, the experimenter drew a ticket from an urn that contained 20 tickets, each of which had either "deck 1" or "deck 2" written on it. Participants were not told which ticket was actually drawn by the experimenter, because that would have indicated the deck from which the card had been drawn. However, they were given the numbers of tickets (out of 20) that had "deck 1" and "deck 2" written on them. For example, in one problem, the urn contained 10 "deck 1" tickets and 10 "deck 2" tickets, so the prior probabilities of the two hypotheses were equal (i.e., both *p*s = .5). We set up problems with three different combinations of prior probabilities: equal (i.e., both *p*s = .5 vs. .5); unequal (*p* = .75 vs. *p* = .25); and extremely unequal (*p* = .95 vs. *p* = .05) (see Appendix A). The combination of the different levels of prior probabilities and the different amounts of information that were transmitted by the present and absent clues resulted in a range of posterior probabilities that ranged from a minimum of *p* = .53 to a maximum of *p* = .98 for the normatively favored hypothesis across all of the problems (see Appendix A). In both studies, we asked participants to make two kinds of judgments; the first was a judgment regarding which of the two hypotheses was the most likely, and the second was a judgment about their confidence regarding the correctness of their first response (see Appendix B). The latter judgment was used to obtain converging evidence about the sources that influenced the probability judgments that the participants made.

Booklets that consisted of a cover page (which was used to collect demographic data from the participants), a set of instructions (which included a sample problem to familiarize the participants with the task), and the 10 problems were distributed individually to students who had been recruited from the University of Milano-Bicocca. The participants were either recruited from quiet study rooms (Study 1) or were tested in a laboratory in exchange for course credit (Study 2). In Study 1, the order in which two alternative conclusions for each problem (i.e., "deck 1" and "deck 2") were presented was fully balanced, which meant that two versions of the questionnaire were created. In Study 2, we balanced both the order of the conclusions and the order in which the 10 problems were presented by presenting them in opposite orders in two versions of the questionnaire; this resulted in the use of a total of four parallel versions of the questionnaire.

Balancing of the present vs. absent clues in the probability tables

In each problem in each study, both the present and the absent letters were chosen randomly from the set {B, C, D, F} (see Appendix A). This procedure was followed to prevent the participants from learning that the subsets of the present and absent clues were always located in the same part of the probability table; that discovery could have systematically oriented their attention.

Another possible source of influence on the judgments that were made by the participants that we considered was the informativeness of the clues (which was computed as IG). The mean IG that was associated with the subset of present clues in the 10 problems (M = .09, SD = .46) was not significantly different from the mean IG that was associated with the subset of absent clues (M = .13, SD = .50), $t(9) = -.19$, $p = .852$, two-tailed. We then considered the IG that was conveyed by each clue separately (the IG was computed for each clue independently of any other clue and treating each clue as if it were present). We found that, in the 10 problems, the mean IG of the 20 present clues (M = .09, SD = .37) was not significantly different from the mean IG of the 20 absent clues (M = .05, SD = .24), $t(19) = .64$, $p = .529$, two-tailed. Furthermore, the IG was greater than .5381 bits (which was the maximum value of the IG among the absent clues) for only 3 of the 20 present clues in the overall probability table (see Appendix A). In contrast, none of the 20 absent

clues had an IG value that was greater than the maximum IG that was transmitted by a present clue, .6349 bits. Finally, the single clue that had the highest IG was an absent clue in only 2 of the 10 problems; in the remaining 8 problems, the most informative clue was present on the drawn card. However, the frequency with which a negative IG occurred (i.e., greater uncertainty about the outcomes after viewing the clue than prior to receiving it) was equally distributed: seven present clues and seven absent clues had negative IG values.

We concluded that, in general, our stimuli were sufficiently balanced to ensure that an overrating of the present clues relative to the absent clues would not be due to a simple preference for the most informative clues.

Study 1

In the first study, we tested the each participant's ability to use base-rate information and incoming evidence (i.e., both present and absent clues) to make discrete judgments regarding the plausibilities of two mutually exclusive and jointly exhaustive hypotheses.

Method

Participants

A total of 40 students (25 females, 15 males; mean age = 21.1 years; range: 19-27 years) from the University of Milano-Bicocca volunteered to participate in the study.

Materials and procedure

The materials and procedure in this study were as described in the Materials and Procedure subsection of the Overview of the Studies section. In Study 1, we asked each participant to make a discrete prediction about which of the two hypotheses (i.e., deck 1 vs. deck 2) was more likely. Furthermore, each participant was asked to rate his/her confidence in his/her discrete prediction on a scale from 1 (*not very confident*) to 7 (*very confident*) (see Appendix B).

Data analysis

Data analyses were conducted using mixed-effects multiple regression models (Bates & Maechler 2009; Pinheiro & Bates, 2000) with random intercepts for the subjects. This approach was

chosen due to a number of theoretical considerations. First, probability is a continuous variable by nature, so it is more appropriate to model it as a continuous predictor than to treat it as a dichotomous factor that indicates the normative-favored deck (Baayen, 2010; DeCoster, Iselin, & Gallucci, 2009). Conveniently, regression models allow us to investigate the ways in which differences in probabilities affect the choices that participants make. Second, it is likely that there will be a certain amount of variability among the answers that are given by the participants in this type of task (e.g., Gigerenzer & Hoffrage, 1995, Table 3, p. 695, Table 4, p. 697; Villejoubert & Mandel, 2002, Figure 3, p. 175); therefore, we introduced a random effect of subject into the models to account for this variability and to obtain a more reliable estimation of the fixed effects. Third, this type of modeling permits a coding scheme for the dependent variable that is not as theory driven as the coding scheme that is usually adopted. In fact, as described below, this type of modeling allowed us to code each dichotomous answer as simply "deck 1" versus "deck 2" rather than "correct" versus "incorrect." In this way, we avoided coding responses according to an assumption that correctness is fundamentally linked to a specific normative criterion (for a review of several alternative Bayesian models, see Crupi et al., 2007; Nelson, 2005, 2008; Nelson et al., 2010).

In the analyses, the prior probabilities (*PriorProb*), the weight of evidence[2] that was computed for all of the clues (*WE_allClues*), and the weight of evidence that was computed for only the subset of present clues (*WE_presClues*) were introduced as potentially significant predictors (Appendix A reports the exact values that were used to fit the model for each of the predictors that we considered). Both first-level effects and interactions were tested. The analysis started with a full factorial model, which was progressively simplified by removing the effects that did not significantly contribute to the goodness of fit of the model. Model parameters were considered one by one, and they were removed when the result of the likelihood ratio test that compared the goodness of fit of the model before removing the effect of each parameter with the goodness of fit of the model after removing the effect of each parameter was not significant. The random-effects

structure initially included the effect of the participants on the intercept. Random effects of the participants on the predictors (random slopes) were also tested to evaluate whether their inclusion significantly increased the goodness of fit of the model. The inclusion of a random slope would indicate that a given effect varies considerably between subjects and would indicate the presence of potential individual differences.

In a first analysis, the discrete judgments that were made by the participants were used as the dependent variable. A logistic mixed-effects model (Jaeger, 2008) was adopted, in which "deck 1" was used as the reference level. In other words, the predictors in this model were calculated relative to the properties of "deck 2" (control level); the model tested the significances of the various predictors in predicting the odds of choosing "deck 2" in comparison to "deck 1"[3]. A second analysis introduced the degree to which each participant was confident in his/her judgment as the dependent variable. In this case, the predictors were computed based on using the chosen deck as a reference. For each problem, the prior probabilities and weight-of-evidence measures of the chosen deck were introduced as independent variables. The statistical significances of the fixed effects in this analysis were evaluated using a Markov chain Monte Carlo (MCMC) sampling algorithm with 10,000 samples.

Results

The final model that resulted from the analysis of the discrete decisions that the participants had made included the *PriorProb* and the *WE_presClues* as fixed predictors. The *WE_allClues* variable was removed from the model because it did not significantly improve the overall goodness of fit. Table 1 reports the fixed effects that were included. The higher the prior probability and the weight of evidence (considering only the subset of present clues) of "deck 2" were, the more likely it was that participants would choose it. Conversely, the lower the prior probability and the weight of evidence (of present clues only) of "deck 2" were, the more likely it was that "deck 1" would be chosen. No significant interaction emerged, which indicated that the reported effects were additive in nature. The inclusion of random slopes of participants on *WE_presClues* (s.d.=.06) was

necessary, which suggests that the subjects differed in their general degrees of sensitivity to the informativeness of the present clues.

The analysis of the degree to which the participants were confident in their judgments revealed the same set of significant effects (see Table 2). For both the *PriorProb* and the *WE_presClues* variables, the higher the predictor values were, the more confidence the subjects had in their responses. However, random slopes of participants on both the *PriorProb* (s.d.=.73) and the *WE_presClues* (s.d.=.03) variables were included in this model, which indicates that both of these effects varied across participants when a confidence report was requested.

Discussion

Our findings provide evidence that people can make judgments based on base-rate considerations, but that they do not integrate the base rates with all of the new evidence that they obtain according to the multiplicative rule from Bayes' theorem. Instead, their integration patterns are influenced by a psychologically compelling feature of the new information, namely by the presence of the clues, as opposed to their absence. Furthermore, they combine this information with base rates in an additive manner.

The effect of prior probabilities is consistent with results from previous studies that have shown that participants are sensitive to base-rate information when the priors are manipulated within subjects (e.g., Fischhoff et al., 1979). Furthermore, this kind of effect is in keeping with the studies that argue that using sample spaces of random events that are clearly defined (conditions met by card games like ours) sensitize participants to base rates (e.g., Ginossar & Trope, 1987, Experiment 6, pp. 471-472; Koehler, 1996; Nisbett, Krantz, Jepson, & Kunda, 1983).

The feature-positive effect when evaluating two competing hypotheses that we observed is consistent with a recent experimental work (Cherubini et al., in revision), but it is at odds with the results of Christensen-Szalanski and Bushyhead's (1981) study, in which physicians used the absence of a symptom as efficiently as the presence of a symptom when estimating the predictive value of the symptoms for a diagnosis. This inconsistency might be traced back to the realism of

their study. Specifically, the participants in their study might have noted the absence of symptoms because they co-occurred with the presence of other important symptoms (see Christensen-Szalanski and Bushyhead, 1981, p. 934). Furthermore, because the participants in that study were expert physicians, their ability to perform a medical diagnosis task might be greater than the ability of non-expert participants who were faced with an abstract task (e.g., Cherubini, Russo, Rusconi, D'Addario, & Boccuti, 2009).

The formal structure of our problems ensured that the informativeness of the present clues was comparable to that of the absent clues, so we were able to avoid the kind of spurious co-occurrence that is described above. We did not insert the *WE* of the subset of absent clues into the model as a potentially significant predictor because it had a strong negative correlation with the *WE_presClues* (r = -.9). This result is not surprising since we designed the problems so that the hypothesis that was favored by the subset of present clues was always the opposite of the hypothesis that was supported by the subset of absent clues.

The results of the analysis of the dichotomous dependent variable were further corroborated by the analysis of confidence ratings. Specifically, both the *PriorProb* and the *WE_presClues* variables significantly predicted the degree to which participants trusted the responses that they gave. We also found significant random effects which indicated that participants varied in the degree to which they were sensitive to the amount of information that was conveyed by the presence of a group of clues (in both dichotomous judgments and in confidence ratings) and to the prior probabilities (only in confidence ratings). This result is consistent with previous research about probability judgments that reported that there were individual differences in responding to this type of task (e.g., Gigerenzer & Hoffrage, 1995; Villejoubert & Mandel, 2002).

Study 2

In Study 2, we aimed to assess the soundness of the results of Study 1 by adding a more finely tuned judgment to the discrete prediction about which of the two hypotheses was the more likely one (see Appendix B). It has been suggested that the response mode can affect the degree to

which people are sensitive to base-rate information. In one study, Manis, Dovalina, Avis, & Cardoze (1980) hypothesized that a discrete judgment, like the one in Study 1 of the present work (or other discrete judgments that participants make in studies of probability matching), might "involve a more rudimentary form of cognitive processing" (Manis et al., 1980) compared with the kind of processing that is entailed in making a subjective probability judgment. It stands to reason, therefore, that the base-rate information would have a more profound effect on discrete predictions compared with the continuous judgments that people make. Although the results from Manis et al.'s study (1980) were reinterpreted by Bar-Hillel and Fischhoff (1981) in light of a possible artifact of the materials they used, their hypothesis remains a reasonable one.

Method

Participants

A total of 48 participants (40 females, 8 males; mean age = 23.1 years; range: 19-48 years) from the University of Milano-Bicocca took part in the study, most of them in exchange for course credit.

Materials and procedure

The materials and procedure were the same as those that have been described in the Materials and Procedure subsection of the Overview of the Studies section. The main difference between Study 1 and the present experiment was that we also asked participants in Study 2 to make a more nuanced judgment about the more likely hypothesis. Specifically, we first asked participants to provide a discrete judgment about the more likely hypothesis, after which we asked them to indicate the percentage between 51% and 100% on an 11-point scale that corresponded to the level of plausibility of the chosen hypothesis (see Appendix B).

Data analysis

The statistical procedures that were described for the first study were also followed in Study 2. The same analysis that had been employed to analyze data regarding the confidence levels of the

participants was used to analyze the data about the newly introduced response (the judgments that the participants made regarding the probability that the chosen deck was the more likely).

Results

The analyses of the discrete preferences of the participants confirmed the results of Study 1. As shown in Table 3, the final model included both the *PriorProb* and the *WE_presClues* variables as significant positive predictors and random slopes of participants on the *WE_presClues* (s.d.=.09). The results of the model of participants' confidence were not completely consistent with the results of Study 1; a significant correlation between the *PriorProb* and *WE_presClues* variables emerged, but no random slopes were found (Table 4). The complex interaction (Figure 1) indicates that both a high prior probability and a high *WE* (computed based on only the present clues) of the chosen deck are necessary for participants to report high levels of confidence in their responses.

In the model that was dedicated to the continuous judgments that the participants made, both the *PriorProb* and the *WE_presClues* emerged as significant predictors (Table 5). The higher the prior probability of the chosen deck, the higher participants perceived its likelihood of being the deck from which the card was drawn. The same reasoning applies to the *WE_presClues*. The likelihood that the drawn card had originated from the chosen deck was deemed to be higher for high levels of the *WE_presClues* that was associated with the chosen deck. No significant interaction emerged, which confirms the additive nature of the two effects. Again, there was no significant impact associated with the *WE_allClues* variable, so it was removed from the model. No random slopes were found.

Discussion

The results of Study 2 are highly consistent with those of Study 1, which provides evidence for the reliability of the observed effects. We replicated the finding that the influences of base-rate information and new information (only if it was about the present features) were additive when considering the discrete dependent variable (i.e., the dichotomous choice between deck 1 and deck 2). The additional dependent variable, a subjective probability judgment about the degree of

plausibility of the chosen deck, yielded the same fixed effects. Hence, our findings do not support Manis et al.'s (1980) hypothesis that the response mode has an effect on the use of prior probabilities. Indeed, base rates affected the judgments that the participants made whether these judgments were dichotomous or were made along a more graduated scale.

The only divergent finding between the present experiment and the results of Study 1 is the significant interaction between the *PriorProb* and the *WE_presClues* variables that emerged from the analysis of the degrees to which the participants were confident in their responses. Participants trusted their judgments more when both of these sources of influence attained high values, and their confidence in their responses decreased when one of these two sources only attained a low value (see Figure 1). This inconsistency can be explained by the different phrasing of the question in the two studies. In Study 2, the question concerned the participant's confidence in both the discrete and the graduated judgments, whereas in Study 1, it only concerned the discrete judgment (see Appendix B). Therefore, the participants had to be supported by high values of both the *PriorProb* and the *WE_presClues* variables to be confident in both of the responses they gave.

General Discussion

In the two studies that are presented in this article, we found consistent evidence that the judgments that the participants made regarding which of two hypotheses was more likely were affected by both prior probabilities and by new information (that was gleaned solely from present clues), but not by an interaction between these variables. This tendency to "weight and add the cues" (Juslin, Nilsson, & Winman, 2009) has long been known to affect belief-updating tasks (e.g., Hogarth & Einhorn, 1992; Juslin et al., 2009; Lopes, 1985, 1987; McKenzie, 1994; Shanteau 1970, 1972, 1975). Our findings extend this understanding to a one-shot hypothesis-testing task by showing that participants integrated prior probabilities and the new information additively instead of relying on multiplicative (Bayesian) integration (e.g., Juslin et al., 2009). However, this result also extends the existing literature by showing that additive integration might involve only *part* of the new information, namely the features that occur relative to those that do not (i.e., the feature-

positive effect). The analysis of the degrees to which participants were confident in their responses provided converging evidence for the additive roles of prior probabilities and the presence of information. However, in Study 2 we found an interaction between these two sources of information that may reflect the different task requirement.

The generalization of the use of additivity as an aggregation rule to one-shot tasks is relevant because this rationale has only been applied and simulated rarely with respect to one-shot tasks (Juslin et al., 2009; McKenzie, 1994). In fact, McKenzie (1994) noted: "Although researchers have examined the conditions under which subjects do and do not use base rates in one-shot tasks, how the information is integrated into a final response is unclear". He then suggested that base rates might be used and averaged with the information that is provided by the LR. These kinds of averaging strategies performed well relative to the normative Bayesian criterion in simulations. The present research represents a tentative answer to the question of whether and when people actually use intuitive strategies, such as averaging strategies, in one-shot tasks. In particular, we found that instead of averaging, the participants simply added the two sources of probabilistic information that they were given in a laboratory card game.

This finding suggests that additivity might not be due to the sequential nature of a belief-updating task; it can also occur when a single revision process is required. The use of additivity as a "composition" rule (Slovic & Lichtenstein, 1971, p. 661) has been considered an error in the combining process, and it has been called "*misaggregation*" in previous works (Fischhoff & Beyth-Marom, 1983, p. 248). The direct consequence of the failure to aggregate the priors and the LR multiplicatively is that people provide probability estimates that are less extreme than they would be if they were acting according to Bayes' rule. Overall, this implies a relatively low certainty (low positive IG) about the hypotheses after receiving new evidence and a loss of information in the case of extreme initial beliefs (i.e., negative IG). In short, the effect of multiplying the priors and the LR is more pronounced than the effect of combining them additively. Future studies should investigate the implications of additivity in one-shot tasks for human-environment relations despite the fact that

this composition rule fails to adhere to the formal rules of probability theory (Juslin et al., 2009; Koehler, 1996; McKenzie, 1994).

Our study also adds to the existing information integration literature by shedding light on one of the possible relevant "weighting parameters" (Slovic & Lichtenstein, 1971, p. 661). We found that the new information influenced the judgments that the participants made only when it gave clues about presence. Linear relationships between participants' judgments and both prior probabilities and the evidential strength of the subset of present clues were observed when the participants made a dichotomous choice between the two competing hypotheses (Studies 1-2) and when they were asked to make a more specific, graduated judgment (Study 2). Thus, participants failed to perceive the informativeness of all of the evidence that they received (i.e., "*misperception*", Fischhoff & Beyth-Marom, 1983, p. 248) because their decision-making processes were guided by a feature that was psychologically relevant but logically irrelevant: the presence of a clue, as opposed to the absence of one.

In conclusion, this is the first study to demonstrate the difficulties in achieving a multiplicative (Bayesian) aggregation of base rates and the LR that might emerge as a consequence of both the misaggregation and misperception of available information sources in an unfamiliar one-shot task of hypothesis testing. Not only do people deviate from the normative Bayesian criterion when assessing the component probabilities by giving more weight to the presence of features than they do to the absence of them, but people may also use a non-normative, additive rule for combining the base rates with the LR to reach a final decision.

References

Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *Journal of Personality and Social Psychology, 35*, 303–314. doi:10.1037/0022-3514.35.5.303

Baayen, R.H. (2010). A real experiment is a factorial experiment? *The Mental Lexicon, 5*, 149-157. doi:10.1075/ml.5.1.06baa

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*, 211–233. doi:10.1016/0001-6918(80)90046-3

Bar-Hillel, M., & Fischhoff, B. (1981). When do base rates affect predictions? *Journal of Personality and Social Psychology, 41*, 671–680. doi:10.1037//0022-3514.41.4.671

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

Bates, D., & Maechler, M. (2009). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-31. http://CRAN.R-project.org/package=lme4

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

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

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

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

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

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

DeCoster, J., Iselin, A. R., & Gallucci, M. (2009). A conceptual and empirical examination of justifications for dichotomization. *Psychological Methods, 14*, 349-366. doi:10.1037/a0016956

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

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

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational behavior and human performance, 23*, 339–359. doi:10.1016/0030-5073(79)90002-3

Fox, C. R., & Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science, 14*, 195–200. doi:10.1111/1467-9280.02431

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

Ginosar, Z., & Trope, Y. (1980). The effects of base rates and individuating information on judgments about another person. *Journal of Experimental Social Psychology, 16*, 228–242. doi:10.1016/0022-1031(80)90066-9

Ginossar, Z., & Trope, Y. (1987). Problem solving in judgment under uncertainty. *Journal of Personality and Social Psychology, 52,* 464–474. doi:10.1037//0022-3514.52.3.464

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

Hogarth, R. M. & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*, 1–55. doi:10.1016/0010-0285(92)90002-J

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

Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review, 116*, 856–874. doi:10.1037/a0016979

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

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

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19*, 1–53. doi:10.1017/S0140525X00041157

Lopes, L. L. (1985). Averaging rules and adjustment processes in Bayesian inference. *Bulletin of the Psychonomic Society, 23*, 509–512.

Lopes, L. L. (1987). Procedural debiasing. *Acta Psychologica, 64*, 167–185. doi:10.1016/0001-6918(87)90005-9

Lyon, D. & Slovic, P. (1976). Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica, 40*, 287–298. doi:10.1016/0001-6918(76)90032-9

Manis, M., Dovalina, I., Avis, N. E., & Cardoze, S. (1980). Base rates can affect individual predictions. *Journal of Personality and Social Psychology, 38*, 231–248. doi:10.1037//0022-3514.38.2.231

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. doi:10.3758/BF03193581

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

Neisser, U. (1963). Decision-time without reaction-time: Experiments in visual scanning. *The American Journal of Psychology, 76*, 376–385. doi:10.2307/1419778

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

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

Nisbett, R. E., & Borgida, E. (1975). Attribution and the psychology of prediction. *Journal of Personality and Social Psychology, 32*, 932–943. doi:10.1037//0022-3514.32.5.932

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90*, 339–363. doi:10.1037//0033-295X.90.4.339

Pinheiro, J.C., & Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag.

Poletiek, F. [H.] (2001). *Hypothesis-testing behaviour*. Hove, U.K.: Psychology Press.

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

Shanteau, J. C. (1970). An additive model for sequential decision making. *Journal of Experimental Psychology, 85*, 181–191. doi:10.1037/h0029552

Shanteau. J. (1972). Descriptive versus normative models of sequential inference judgment. *Journal of Experimental Psychology, 93*, 63–68. doi:10.1037/h0032509

Shanteau, J. (1975). Averaging versus multiplying combination rules of inference judgment. *Acta Psychologica, 39*, 83–89. doi:10.1016/0001-6918(75)90023-2

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

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

Slowiaczek, L. M., Klayman, J., Sherman, S. J., & Skov, R. B. (1992). Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory & Cognition, 20*, 392–405. doi:10.3758/BF03210923

Treisman, A. & Souther, J. (1985). Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General, 114*, 285–310. doi:10.1037//0096-3445.114.3.285

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

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.

Footnotes

[1] Note that here we refer to "one-shot tasks" (McKenzie, 1994), in which participants are given base rates and make a single judgment. Participants in "belief-updating tasks" (McKenzie, 1994) make multiple judgments, and the first response corresponds to the prior probabilities for the second revision (e.g., Beach, 1968), so it is more likely that unequal prior probabilities will be available to participants in the latter kind of task.

[2] We adopted *WE* because it is a direct measure of the informativeness of a datum independent of the prior probabilities and of the posterior probabilities, unlike measures such as IG (see the Introduction).

[3] The opposite model, in which "deck 2" was used as the reference level, is statistically equivalent and leads to consistent results.

Table 1

Analysis of participants' discrete judgments: Fixed effects resulting from a logistic mixed-effects regression model ("deck 1" is the reference level). Study 1.

	Estimate	Std. Error	z-value	p-value
Intercept	-1.45	.27	5.49	<.001
PriorProp	3.13	.46	6.79	<.001
WE_presClues	.11	.02	7.08	<.001

Table 2

Analysis of participants' confidence: Fixed effects resulting from a mixed-effects regression model. Study 1.

	Estimate	Std. Error	MCMCmean	pMCMC
Intercept	3.06	.23	3.06	<.001
PriorProb	1.01	.19	1.05	<.001
WE_presClues	.03	.01	.02	<.01

Table 3

Analysis of participants' discrete judgments: Fixed effects resulting from a logistic mixed-effects regression model ("deck 1" is the reference level). Study 2.

	Estimate	Std. Error	z-value	p-value
Intercept	-2.51	.29	8.61	<.001
PriorProb	5.05	.52	9.69	<.001
WE_presClues	.11	.02	6.05	<.001

Table 4

Analysis of participants' confidence: Fixed effects resulting from a mixed-effects regression model. Study 2.

	Estimate	Std. Error	MCMCmean	pMCMC
Intercept	3.95	.19	3.96	<.001
PriorProb	.33	.16	.33	.06
WE_presClues	-.01	.01	-.01	n.s.
*PriorProb * WE_presClues*	.03	.01	.03	<.05

Table 5

Analysis of participants' probability judgments: Fixed effects resulting from a mixed-effects regression model. Study 2.

	Estimate	Std. Error	MCMCmean	pMCMC
Intercept	.67	.02.	.67	<.001
PriorProb	.10	.02	.10	<.001
WE_presClues	.002	.001	.002	<.001

Figure caption

Figure 1. Analysis of participants' confidence in Study 2: Participants' confidence in the responses they gave as a function of prior probabilities and weight of evidence of the subset of present clues only.

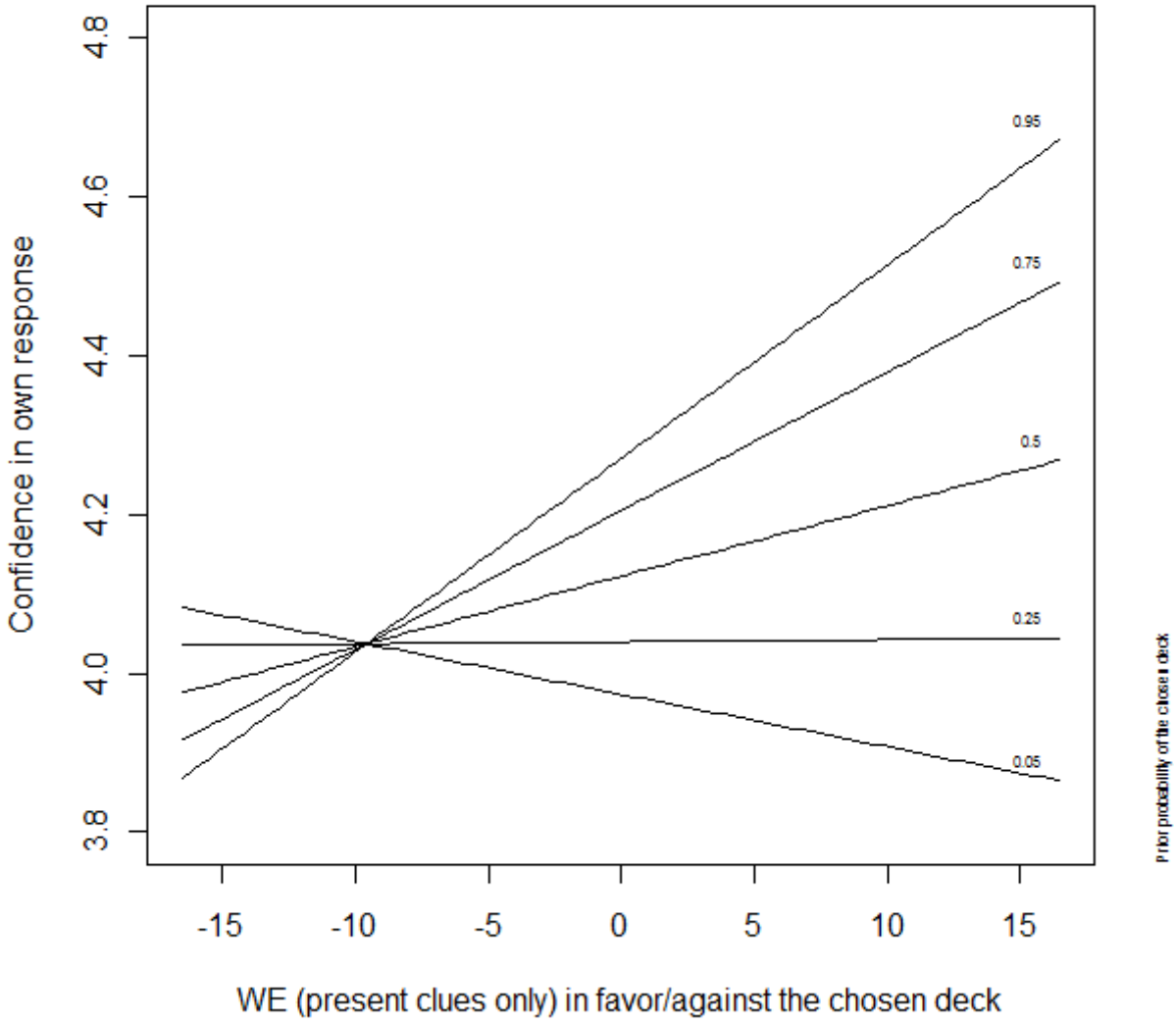

Confidence in own response
4.8
4.6
4.4
4.2
4.0
3.8
0.95
0.75
0.5
0.25
0.05
Prior probability of the chosen deck
-15
-10
-5
0
5
10
15
WE (present clues only) in favor/against the chosen deck

Figure 1

Appendices

Appendix A

Formal properties of the 10 problems that were used in the two experiments. The conditional probabilities of the present clues are presented in bold. The conditional probabilities of the absent clues correspond to the frequencies that were shown to the participants in the tables.

Problem	Deck	Prior probabilities [*PriorProb*]	Likelihoods				Information gain of present clues	Information gain of absent clues	Weight of evidence of present clues in relation to deck 2 (control level) [*WE_presClues*]	Weight of evidence of all clues in relation to deck 2 (control level) [*WE_allClues*]	Posterior probabilities
			p (B)	*p* (C)	*p* (D)	*p* (F)					
1	1	.2500	.6700	**.3300**	**.6900**	.9000	-.1880	.6204	-5.0488	5.3830	.0900
	2	.7500	.5500	**.0800**	**.8900**	.1900					.9100
2	1	.5000	**.5000**	**.1800**	.3500	.1300	.0016	.2592	.4139	-5.3495	.7700
	2	.5000	**.1100**	**.9000**	.7000	.5000					.2300
3	1	.2500	.1300	.3000	**.0900**	**.7300**	.4865	.2218	7.2301	-5.3526	.5300
	2	.7500	.8800	.7200	**.6200**	**.5600**					.4700
4	1	.9500	.7100	**.2300**	**.5800**	.9000	.2724	-.7136	-16.2001	-3.3422	.9800
	2	.0500	.3000	**.0200**	**.1600**	.2000					.0200
5	1	.7500	**.4000**	**.9000**	.7200	.9500	.7170	-.0858	-14.3573	-6.1807	.9300
	2	.2500	**.0400**	**.3300**	.5400	.8000					.0700
6	1	.0500	.1200	.2800	**.0700**	**.5700**	.2275	-.6958	8.8404	-5.3217	.1500
	2	.9500	.9100	.7300	**.4700**	**.6500**					.8500
7	1	.5000	.8500	**.8000**	**.2000**	.7300	.2765	.6437	-5.9989	5.4084	.2200
	2	.5000	.3000	**.0600**	**.6700**	.2000					.7800
8	1	.9500	**.0500**	**.5600**	.6000	.4600	-.7122	.2144	13.1702	5.3887	.8500
	2	.0500	**.7000**	**.8300**	.9400	.4000					.1500
9	1	.7500	.4000	.5000	**.0200**	**.2200**	.3971	.5931	15.1792	5.5268	.4600
	2	.2500	.8700	.7500	**.2900**	**.5000**					.5400
10	1	.0500	.9600	**.8700**	**.9000**	.8900	-.5940	.2817	-16.4825	5.1441	.0200
	2	.9500	.2100	**.0400**	**.4400**	.1900					.9800

Appendix B

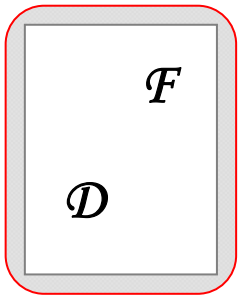
A sample stimulus.

In this case, the urn used to choose the deck from which to draw the card contained 15 tickets with "deck 2" written on them and 5 tickets with "deck 1".

On the card drawn from the chosen deck, there are a D and an F, but not a B or a C.

Mark the box indicating the deck from which the card was more likely drawn.

	B	C	D	F
deck 1	13	30	9	73
deck 2	88	72	62	56



deck 1 deck 2

Indicate your confidence in the response you provided:

not very confident 1 2 3 4 5 6 7 very confident

Study 2 version.

From which deck was the card most likely drawn?

(mark one of the two boxes with an X)

deck 1 deck 2

How likely is it that the card was drawn from this deck? (mark the corresponding percentage)

51% 55% 60% 65% 70% 75% 80% 85% 90% 95% 100%

Indicate your confidence in the responses you provided:

not very confident 1 2 3 4 5 6 7 very confident