Journal of Experimental Psychology:
Learning, Memory, and Cognition
2003, Vol. 29, No. 1, 80–96

0278-7393/03/$12.00 DOI: 10.1037/0278-7393.29.1.80

An Associative Framework for Probability Judgment: An Application to Biases

Pedro L. Cobos and Julián Almaraz
Universidad de Málaga

Juan A. García-Madruga
Universidad Nacional de Educación a Distancia

Three experiments show that understanding of biases in probability judgment can be improved by extending the application of the associative-learning framework. In Experiment 1, the authors used M. A. Gluck and G. H. Bower's (1988a) diagnostic-learning task to replicate apparent base-rate neglect and to induce the conjunction fallacy in a later judgment phase as a by-product of the conversion bias. In Experiment 2, the authors found stronger evidence of the conversion bias with the same learning task. In Experiment 3, the authors changed the diagnostic-learning task to induce some conjunction fallacies that were not based on the conversion bias. The authors show that the conjunction fallacies obtained in Experiment 3 can be explained by adding an averaging component to M. A. Gluck and G. H. Bower's model.

A great deal of research in decision making has focused on people's biases in probability judgment. The main reason for this is that deviations from normative theories are more informative than correct estimations in order to infer the cognitive processes underlying probability judgment, which, in turn, is a basic component of decision making. Some of these biases are remarkable because they seem to involve the violation of some basic and simple principles of the probability theory.

Our main purpose is to show that the associative-learning framework can improve our understanding of biases in probability judgment. We focus on three well-known errors: the base-rate neglect, the conjunction fallacy, and the conversion of conditional probabilities (thereafter referred to as the conversion bias). First, let us briefly outline these biases.

Some Biases in Probability Judgment

The Base-Rate Neglect

The base-rate neglect has been found in problems with a Bayesian structure. Basically, participants' task is to estimate the probability of an object or person belonging to a series of mutually exclusive categories. From Bayes's theorem, the following formula was established to calculate the probability of the object belonging to one of the categories (e.g., category *A*)

$$p(A|O) = \frac{p(O \mid A) \times p(A)}{p(O \mid A) \times p(A) + p(O \mid \bar{A}) \times p(\bar{A})},$$

where *O* is the object to be classified, $p(A)$ and $p(\bar{A})$ are the base rates of the target category and of its complementary set, respectively, $p(A \mid O)$ is the probability that object *O* belongs to category *A*, also called the inverse probability, and $p(O \mid A)$ and $p(O \mid \bar{A})$ are the probabilities of object *O* conditional on category *A* and conditional on category $\bar{A}$, respectively (i.e., the conditional probabilities or diagnostic information).

The base-rate neglect is obtained when participants' judgments are not affected, or are only slightly affected, by the category frequencies or the base rates. This phenomenon has been mainly studied in cases in which there is some kind of conflict between the base rates and the diagnostic information. Specifically, when $p(O \mid A) > p(O \mid \bar{A})$, and $p(A) < p(\bar{A})$. In such cases, participants' judgments are more closely correlated with the diagnostic information than with the base rates (Kahneman & Tversky, 1973). In other words, participants tend to favor the less-frequent category.

The Conjunction Fallacy

The conjunction fallacy occurs when the probability judgment for a conjunction of events is higher than for one of its constituents. This implies the violation of the conjunction theorem that states that the probability of a conjunction of events is always lower or equal to the probability of any of the events that constitute the conjunction, that is $p(A \cap B) \leq p(A)$, and $p(A \cap B) \leq p(B)$. In the *Linda* problem (Tversky & Kahneman, 1983), the most cited in this context, a character named Linda is described to participants as a female activist. Later, participants have to sort a series of statements in terms of their probability in a decreasing order. The

Pedro L. Cobos and Julián Almaraz, Departamento de Psicología Básica, Universidad de Málaga, Málaga, Spain; Juan A. García-Madruga, Departmento de Psicología, Universidad Nacional de Educación a Distancia, Madrid, Spain.

Juan A. García-Madruga is now at the Departmento de Psicología Evolutiva y de la Educación, Madrid, Spain.

This work is part of a research project supported by Junta de Andalucía (HUM0105) and Ministerio de Ciencia y Tecnologia (BSO2000-1216). We are very grateful to Francisco J. López and David R. Shanks for their valuable comments on drafts of this article. We also thank the reviewers' outstanding work on previous versions.

Correspondence concerning this article should be addressed to Pedro L. Cobos, Departamento de Psicología Básica, Facultad de Psicología, Universidad de Málaga, Campus de Teatinos, Málaga 29071, Spain. E-mail: p_cobos@uma.es

most important statements are: *Linda is active in the feminist movement* (A), *Linda is a bank teller* (B), *Linda is a bank teller and is active in the feminist movement* (A and B). As a result, many participants estimate that statement A and B is more likely than statement B (e.g., Agnoli & Krantz, 1989; Fiedler, 1988; Politzer & Noveck, 1991; Wolford & Taylor, 1990).

The Conversion Bias

The conversion bias has been found in Bayesian problems, and has been studied in relation to the base-rate neglect. As we stated previously, in Bayesian problems, participants receive information about category base rates and conditional probabilities [e.g., $p(O|A)$], and they have to estimate the inverse probability [$p(A|O)$]. The conversion bias is conceived as a tendency to equate the inverse probability with its corresponding conditional probability [i.e., $p(A|O) = p(O|A)$]. For example, Braine, Connell, Freitag, and O'Brien (1990) found some cases of conversion bias using a version of Kahneman and Tversky's (1982) *taxicab problem*.[1] One of the consequences of the conversion bias is that judgments are insensitive to base rates. This is why some researchers conceive the conversion bias as the cause of the base-rate neglect (Braine et al., 1990, Dougherty, Gettys, & Ogden, 1999; Fiedler, Brinkmann, Betsch, & Wild, 2000; Gavanski & Hui, 1992; McMullen, Fazio, & Gavanski, 1997; Sherman, McMullen, & Gavanski, 1992).

The Associative Learning Framework (AL)

The more lasting and widely accepted theories in the field to explain biases in probability judgment have been based on the cognitive-illusion hypothesis (Kahneman & Tversky, 1973, 1996). According to this hypothesis, the experimental tasks contain some elements that elicit certain heuristics based on processes of natural assessment that are appropriate for some tasks in the extraexperimental context but not for tasks in the experimental one. Such processes are effective to estimate among other things, the degree of association between events (objects, or features), the consistency of given information, the similarity between objects, or the representativeness of objects regarding a given category. As these processes were conceived as rapid, effortless, and based on the properties of objects (instead of the extensions of sets), they were taken as the basis for intuitive reasoning, as opposed to extensional reasoning, which constitutes the basis for the probability theory (Tversky & Kahneman, 1983).

During the past few years, these theories have been criticized for their minimal quantitative accounts (i.e., the specification of the computational processes underlying probability judgment). This dearth is considered the reason for the slight progress in understanding the cognitive processes involved in probability judgment biases, despite more than 30 years of research (see Gigerenzer, 1996, but see Kahneman & Tversky, 1996, for an alternative viewpoint).

In the present article, we extend AL to probability judgment, and, specifically, to the biases above mentioned (see Lagnado & Shanks, 2002; López, Cobos, Caño, & Shanks, 1998; Sloman, 1996; and Windschitl & Weber, 1999, for similar proposals). There are two important reasons for choosing an associative framework. First, associative mechanisms contain the main features attributed to intuitive reasoning (they are rapid, effortless, and their inputs are the properties of objects), and have been proposed as the basis for processes of natural assessments (e.g., degree of association between events or features, between-objects similarity, and representativeness). In fact, some authors have already used the term *intuitive* to characterize the processing carried out by associative systems (Hinton, 1990, Sloman, 1996, Smolensky, 1988). Second, associative mechanisms can be specified in terms of formal processing models. Thus, AL can potentially overcome the lack of quantitative descriptions attributed to the cognitive-illusion theories and, then, can be useful to specify the precise conditions in which biases occur.

According to our suggestion, an associative-learning mechanism underlies probability judgments when the contents of the probability-judgment task form part of a previous experiential-learning situation; that is, a situation where an individual has to learn to predict some consequent events (outcomes) from other antecedent events (cues) on the basis of repeated exposure to their temporal distribution (see Shanks, 1991b, for a definition of *experienced* situations). This repeated exposure favors the formation of associations between the representations of the cues and of the outcomes so that, when a given cue is perceived, the representation of the associated outcome is automatically activated. The amount of activation of the outcome representation reflects the degree of relationship between the cue and the outcome.

The output of the associative mechanism (the activation of the outcome representation) may determine a probability judgment in two different ways either (a) directly or (b) serving as the input for a further heuristic process that, ultimately, produces the response. Eventually, this kind of probability judgment will violate some principle of the experimenter's normative analysis of the probability-judgment task, and thus, a bias will occur.

In addition, the AL framework includes two assumptions that play an important role in explaining probability judgments and biases. First, associative learning is sensitive to the relative validity of cues as predictors of outcomes. That is, the predictive value of a target cue is not an absolute value but relative to the predictive value of other cues with which the target cue co-occurs. There is much empirical evidence showing that the perceived relationship between a cue and an outcome is sensitive to the relative validity of the cue, both in animal as well as in human experiential learning (Baker, Mercier, Vallée-Tourangeau, Frank, & Pan, 1993; Chapman & Robbins, 1990; Cobos, Cano, López, Luque, & Almaraz, 2000; Dickinson & Shanks, 1985; Kamin, 1968; Matute, Arcedi-

[1] In this problem, it is said that taxicabs of different colors, which belong to different companies, operate in a city. The percentage of cabs of each color (the base rates) is as follows: 10% are red, 15% are blue, 20% are yellow, 25% are orange, and 30% are green. A given cab is involved in a hit-and-run accident in the presence of a witness, but because of bad visibility conditions, her testimony was subject to inaccuracies. In one of the problems used by Braine et al. (1990), participants were told that if the cab was blue, then the witness said it was blue 80% of the time (the conditional probability). Then, they had to judge the probability of the cab involved in the accident being blue given that the witness said it was blue (the inverse probability). Sixty-six percent of the ratings were 80%, that is, just the conditional probability; 6% of the ratings were close to 80% (what the authors called fuzzed conversion), and the remaining ratings were distributed in small quantities throughout four more categories.

ano, & Miller, 1996; Shanks, 1991a; Shanks & López, 1996; Wagner, Logan, Haberlandt & Price, 1968). In the context of associative-learning models, this phenomenon is also known as *cue competition* because cues are seen as competing for the limited amount of associative strength supported by outcomes whenever such cues co-occur in the same learning trial. As a consequence, cues with a high predictive value for a certain outcome prevent other redundant, or less valid, cues from gaining associative strength for that outcome.

The second assumption is that associative learning is sensitive to learning order. Specifically, the formation or use of associative links from cues (antecedent events) to outcomes (consequent events) are primed over the formation or use of associative links in the opposite order. This should not be taken as precluding the possibility of associative links in the opposite temporal order. However, there is a good amount of evidence showing that individuals trained in the cue–outcome temporal order find it difficult to go in the outcome–cue direction. For example, classical conditioning is much more easily obtained in conditioned stimulus–unconditioned stimulus preparations than in unconditioned stimulus–conditioned stimulus preparations. In fact almost all classical conditioning models assume unidirectional associations (see, however, Miller & Barnett, 1993, or Gerolin & Matute, 1999). People also find it difficult to make judgments in the outcome–cue order if they have been trained following a cue–outcome temporal order (see Fiedler et al., 2000; Price & Yates, 1995; Waldmann, 1996).

There are many associative models that can be used to implement the assumptions of AL framework. For reasons of economy, we have used a bilayered feed-forward neural network that updates its weights by means of the delta rule, which is formally equivalent to the Rescorla-Wagner (RW) rule for associative learning (Rescorla & Wagner, 1972, Sutton & Barto, 1981). The delta rule is sensitive to the relative validity of cues, and is the most widely used to explain cue-competition effects. For simplicity, we have opted for an elemental representation of compound cues, i.e., each input unit represents a different single cue [see, however, Gluck & Bower's (1988b) configural model, attentional learning covering map (ALCOVE; see Kruschke 1992; Kruschke & Johansen, 1999, for a later version of ALCOVE), Pearce's configural model (Pearce, 1994), or attention to distinctive input (ADIT; Kruschke, 1996)].

Other nonassociative, quantitative accounts have also been proposed since the early 1990s to overcome the shortcomings of the cognitive-illusion theories. The theory of probabilistic mental models (PMM theory; see Gigerenzer, 1993; Gigerenzer & Goldstein, 1996; Gigerenzer, Hoffrage, & Kleinbölting, 1991), the model of Brunswikian induction algorithm for social cognition (BIAS model; see Fiedler, 1996) and the model MINERVA-decision making (MDM model; see Dougherty et al., 1999) are remarkable examples since they have been proposed as general frameworks to account for a wide variety of biases in probability judgment. We will return to these alternative accounts in the *General Discussion* section to better appreciate our contribution.

Research Strategy

An important aspect of our research concerns the strategy adopted to provide support for the AL. The rationale is as follows: If some biases in probability judgment can be caused by associative processes elicited by a previous experiential-learning situation, then the use of an experiential-learning task might induce such biases if participants are required to make probability judgments in a later test. The basic procedure provided participants with information about the relationship between a series of events through an experiential-learning task rather than through verbal descriptions or numeric presentations, as it is usually done. After that, participants were asked to judge the probability of a series of statements about the same events. It will become evident here that an interesting advantage of this strategy, together with the use of quantitative models, is that it allows for studying and specifying the precise conditions in which biases occur.

In the last few years, there has been increasing agreement that the processes underlying probability judgment and decision making are adapted to deal with information acquired from experience (Dougherty et al., 1999; Fiedler, 1996; Fiedler et al., 2000; Gigerenzer & Hoffrage, 1995). However, though this assumption has been taken seriously in the last several years, there are still few studies using learning tasks with unknown contents (i.e., participants have no prior beliefs) to elicit biases. Though this strategy does not substitute for others more commonly used, we think that its use will provide stronger support for quantitative theories such as AL, BIAS, or MDM.

In our experiments, we have used Gluck and Bower's (1988a) diagnostic-learning task and another modified version of such task. Gluck and Bower showed that their task elicited a sort of base-rate neglect, known as *apparent base-rate neglect*, that could be explained by a simple bilayered adaptive-neural network as the one described before. Our experiments may be viewed as an extension of Gluck and Bower's experiments to other biases.

Overview of the Experiments

In Experiment 1, we used Gluck and Bower's (1988a) task to induce the conjunction fallacy. During the training phase, participants had to learn to diagnose diseases from symptoms (i.e., *symptom–disease direction*). The strategy to obtain the conjunction fallacy consisted in requiring conditional-probability judgments in the opposite direction during the test phase, i.e., the probability of symptoms conditional on the diseases. For example, participants had to judge the probability of some conjunctive symptoms as well as the probability of each constituent symptom given one of the diseases. We expected to find some conjunction fallacies as a byproduct of a conversion bias. A secondary objective of Experiment 1 was to replicate Gluck and Bower's apparent base-rate neglect.

Because part of the explanatory power of the conjunction fallacies obtained in Experiment 1 was on the conversion-bias hypothesis, in Experiment 2 we looked for independent and stronger evidence of the latter bias. The same learning task was used in this experiment, though some changes were introduced in the test phase.

In Experiment 3, we explored whether the conjunction fallacies could also be obtained even if the conditional probability judgments in the test phase had a cue–outcome direction. In other words, we looked for conjunction fallacies that are not the byproducts of the conversion bias. This is an important issue because there is some evidence showing that conjunction fallacies are frequently obtained when people make conditional probability judgments consistent with the direction of associative links (Tversky & Kahneman, 1983). By finding such conjunction fallacies we wished to explore the usefulness of experimental-learning tasks to improve our understanding of biases. Though such conjunction fallacies are not directly predicted from the AL, we will show that they are, nonetheless, compatible with the central assumption that associative-learning mechanisms may underlie probability judgments when people are previously engaged in an experiential-learning situation.

Experiment 1

In Experiment 1, we used Gluck and Bower's (1988a) diagnostic-learning task, though with some minimal changes, to induce the conjunction fallacy and, secondarily, to replicate their apparent base-rate neglect. Participants had to learn to diagnose two possible diseases from a series of symptoms. Learners went through a series of 160 trials during each of which they had to diagnose the disease of a hypothetical patient from the symptoms she presented. After the diagnosis, participants received corrective feedback. Patients could exhibit any possible combination of 4 different symptoms (S_1, S_2, S_3, and S_4). This amounted to 16 possible combinations. Seventy-five percent of the patients suffered from a common disease (*C*) while the remaining 25% suffered from a rare disease (*R*). Table 1 shows the programmed probabilities that determined the relationship between each symptom and each disease. Given any of the diseases, the presence of the different symptoms was independent from each other.

As can be seen in Table 1, probabilities were arranged so that $p(R) < p(C)$, but $p(S_1|R) > p(S_1|C)$. Though S_1 was paired with *R* as many times as with *C*, Gluck and Bower (1988a) found that participants judged $p(R|S_1)$ more probable than $p(C|S_1)$, i.e., a case of (apparent) base-rate neglect. However, their experiments revealed that participants' biases toward the rare disease were due to their sensitivity to the relative validity of the symptoms as predictors of the diseases. That is, S_1 was the best predictor of *R* and the worst predictor of *C* (compared with the remaining symptoms). Gluck and Bower also showed that a neural-network model was able to account for their apparent base-rate neglect on the basis of its sensitivity to the relative validity of the symptoms as predictors of the diseases. These results have also been replicated by others (Cobos, López, Rando, Fernández, & Almaraz, 1993; Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Kruschke, 1996; Myers, Lohmeier, & Well, 1994; Nosofsky, Kruschke, & McKinley, 1992; Shanks, 1990).

Table 1
Programmed Probabilities of Each Symptom Given Each Disease and of Each Disease Given Each Symptom for the Diagnostic-Learning Task Used in Experiments 1 and 2

	Diseases							
	C				*R*			
Direction	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4
$p(S\|D)$	.20	.30	.40	.60	.60	.40	.30	.20
$p(D\|S)$	.50	.69	.80	.90	.50	.31	.20	.10

Note. *C* = common; *R* = rare; $p(S|D)$ = programmed probability of the symptom conditional on the disease; $p(D|S)$ = programmed probability of the disease given the symptom; *S* = symptom; *D* = disease.

We extended the use of this experiential-learning task to induce the conjunction fallacy by asking participants to judge the probability of the symptoms (*S*) conditional on the diseases (*D*): $p(S|D)$ judgments. According to the usual strategy to obtain the conjunction fallacy, we combined a symptom that is weakly or inhibitorily associated with the disease with another symptom that is strongly associated with the disease. We then compared the probability judgment for the combined symptoms with that for the weakly associated constituent. For instance, as can be inferred from Table 1, S_1 should have an inhibitory associative strength with *C*, whereas S_4 should have a strong positive-connection strength with *C*. Thus, participants should judge the conjunction S_1 and S_4 as more likely than the constituent S_1 in patients suffering from *C*.

From an AL perspective, the symptom–disease learning order used during the training phase primes the use of symptom–disease directed associations even if participants have to judge the probability of the symptoms conditional on the diseases. Now suppose, for example, that $p(S_4 \mid R)$ and $p(S_4 \text{ and } S_1|R)$ have to be judged. According to the programmed probabilities in Table 1, S_4 and S_1 should acquire an inhibitory and a positive associative strength, respectively, for *R*. Thus, the addition of S_1 should raise the activation of Disease *R* representation with respect to S_4 alone. Therefore, the probability of S_4 and S_1 should be judged greater than the probability of S_4 in patients suffering from *R*. In other words, a conjunction fallacy is expected in this case.

In addition, there are other interesting predictions. For example, in patients suffering from *R*, Symptom S_1 should receive higher probability ratings than the conjunction S_4 and S_1. This follows from the fact that the summed associative strength for *R* should be higher in the first case than in the second case. For the same reason, in patients suffering from *C*, Symptom S_4 should receive higher ratings than the conjunction S_1 and S_4. Using traditional tasks, Tversky and Kahneman (1983) have obtained evidence consistent with these predictions. Finally, another important prediction is that participants' judgments for the conjunction S_1 and S_4 in patients suffering from *C* should be higher than for the same conjunction (S_4 and S_1) in patients suffering from *R*. According to the programmed probabilities (see Table 1), there should be no difference between those judgments. However, if participants commit a conversion bias in $p(S|D)$ judgments, such judgments should be sensitive to the disease base rates. The associative basis for the expected difference is that $S_1S_4 \rightarrow C$ trials are more frequent than $S_1S_4 \rightarrow R$ trials. Thus, the summed associative strength of S_1 and S_4 for *C* will be higher than that for *R*. For the same reason, Symptom S_4 and the conjunctions S_1 and S_3 as well as S_1 and S_2 in patients suffering from *C* should receive higher ratings than Symptom S_1 and the conjunctions S_4 and S_2 and S_4 and S_3 in patients suffering from *R*, respectively. Thus, Experiment 1 served to test all these predictions. We also have simulated participants' results with the simple neural network above described to confirm and illustrate all these predictions more precisely.

Another important objective we aimed at in our three experiments was to improve the experimental procedure to ensure that participants' biases are not the effects of some linguistic confusion or some misinterpretation regarding the judgments required through the test phase. In particular, it is very important that participants do not confuse the probability of a symptom given a disease [e.g. $p(S_4|R)$] with the probability of that symptom in isolation given the same disease [e.g. $p(S_4$ and no S_1 and no S_2 and no $S_3|R)$]. In the latter case, assigning a higher probability to a conjunction than to one of its constituents would not actually be a case of conjunction fallacy. As this is a key issue, we have included three procedural measures that will be specified in the *Procedure* section: instructions, visual aids, and filler items at the test phase.

Another potential problem that has to be avoided is the linguistic confusion between $p(D|S)$ items and $p(S|D)$ items. For instance, if the probability of the conjunction S_1 and S_4 in patients with R is confused with the probability of suffering R in patients with S_1 and S_4, any possible conjunction fallacy could be considered as a mere misinterpretation. Again, in the *Procedure* section we will explain how these problems were avoided.

Method

Participants. Forty psychology undergraduate students from Universidad de La Laguna volunteered to take part in this experiment as partial fulfillment of their course requirements.

Apparatus and stimuli. Each participant stayed in a cubicle in front of a PC-type computer. Stimuli of the training and the test phase were presented on the PC screen, and participants used the keyboard to respond in both phases. The cubicles were not absolutely isolated, although there was a screen separating them to prevent participants from seeing what was taking place at adjoining cubicles.

The fictitious diseases that hypothetical patients could suffer from were *Cajal's disease* and *Vallejo's syndrome* (translated from the Spanish *enfermedad de Cajal* and *sindrome de Vallejo*). The symptoms could be headache, labored breathing, swollen eyes, and loss of balance.

Procedure. First, participants read some instructions and were given additional oral instructions from time to time about the training and the test phases. Such instructions were somewhat long, as they included some explanations and examples to avoid any possible misunderstanding of the probability judgments participants had to make during the test phase. To encourage participants to discriminate between the absence of symptoms and the absence of information about symptoms, they were given the following example:

> Suppose you must judge the probability that a 25-year-old woman will give birth shortly. Now take into consideration a different question: the probability that a 25-year-old woman will give birth shortly, provided that she is not pregnant at present. As you can see with this example, giving no information about a number of elements is completely different from giving specific information about the absence of those elements. In the task you are about to do, when there is information about the absence of a symptom, that symptom appears written in red. On the contrary, when there is no information at all about the presence or absence of a symptom, that symptom appears in white letters and with a question mark on its side.

In addition, participants were given oral instructions and an example in order to help them to distinguish the $p(D|S)$ from the $p(S|D)$ items. In the example, participants were asked to think about the relationship between sneezing and having a cold. The probability of sneezing while having a cold, $p(S|D)$, is close to 1. Conversely, the probability of having a cold if sneezing, $p(D|S)$, is considerably lower if we take into account that sneezing is also a symptom of many other diseases, e.g., an allergy to dust. After this example, as in the previous one, participants were asked if they understood the difference between both types of conditional probability. The instructions were not resumed until every participant had understood it.

As part of the instructions, participants went through a pretraining task to make them familiar with the experimental task and to solve any possible difficulty in both the training phase and the test phase. In this task, the diseases were *coliosis* and *amitismo*, whereas the symptoms were high blood pressure, stiffness, vertigo, and cough. All the types of items included in the test phase of the experimental task were also included in the test phase of this pretraining task.

During the training phase, participants received 160-training trials during each of which the symptoms present in a hypothetical patient were displayed on the screen. Diseases and symptoms were randomly generated according to the programmed probabilities listed in Table 1 on a trial-by-trial basis. Thus, the order in which stimuli were presented, as well as the frequency of symptom patterns, could be different for each participant. Every patient suffered only one of the diseases. On each trial, the four symptoms were listed vertically in the center of the screen in a random order. The symptoms present in a given patient appeared highlighted and in blue letters, whilst symptoms absent appeared in red and were not highlighted. Participants were told the meaning of these visual cues in the instructions. They had to choose between two diseases listed horizontally at the bottom of the screen. To give their diagnoses, participants had to press either Key 1 or Key 2 for each disease. Once the disease was chosen, participants received corrective feedback. If the answer was correct, the word "CORRECT" appeared on the screen. If the answer was wrong, the sentences: "FALSE. The correct answer is *Cajal's disease*/*Vallejo's syndrome*" appeared. These messages remained on the screen for 2 s; then, participants proceeded to the next trial. The roles of the symptoms and diseases shown in Table 1 were randomly assigned for each participant to the actual symptoms and fictitious disease names mentioned above.

The test phase was divided into two different blocks. In Block $p(D|S)$, participants had to estimate the probability of Diseases R and C for each of the 16 possible symptom patterns. They were also asked to estimate the probability of each disease given Symptom S_1 to know whether they tended to choose Disease R over Disease C. To encourage participants to discriminate between the absence of symptoms and the absence of information about symptoms, they had to make judgments on the basis of complete information about the presence or absence of each symptom and judgments on the basis of the presence of a single symptom without any information about the remaining ones. The items in which the probability of Diseases R and C had to be estimated for each symptom pattern (a total of 16) were items with complete information about the presence and absence of symptoms. In addition, there were four more items mixed with the others on which participants only received information about the presence of one symptom. Visual cues were used to help participants interpret the items. Each item consisted of four symptoms listed vertically in the center of the screen. As in the training phase, present symptoms were highlighted and in blue letters while those absent appeared in red. In addition to this, those symptoms about which there was no information available appeared in white letters with a question mark on its right.

In the second block of the test phase, Block $p(S|D)$, participants had to estimate the probability of a series of conjunctive symptoms in patients suffering from Disease C and in patients suffering from Disease R. The probability of conjunctions S_4 and S_1; S_4 and S_2; and S_4 and S_3 had to be estimated in patients with Disease R. Participants had also to estimate the probability of symptoms S_1 through S_4 in patients suffering from R. Thus, we would know in which cases the conjunction fallacy was committed. In patients suffering from Disease C, participants had to estimate the probability of the following conjunctions S_1 and S_4; S_1 and S_3; S_1 and S_2. They were also asked to estimate the probability of Symptoms S_1 through S_4. Thus, the strategy both in the context of C and in the context of R was to compound the least associated symptom with the remaining ones. In addition to this, other filler items were introduced in which participants had to judge the probability of isolated symptoms. Thus, participants had to judge the probability of having only S_1, only S_2, only S_3, and only S_4, conditional on R and on C. These filler items were added to promote the discrimination between the probability of having a symptom and the probability of having a symptom in the absence of the remaining three. Once again, we used the visual cues described above to avoid the misinterpretation of the different items.

To favor the discrimination between $p(D|S)$ and $p(S|D)$ judgments, participants used different scales to make their judgments. The scale for Block $p(D|S)$ was a segmented horizontal line, with Disease R at one end and Disease C at the other. To make their judgments, participants had to place a pointer somewhere along the scale using the arrow keys on the keyboard. Thus, it was made evident that the probability of Disease R was always complementary to that of Disease C. At odds with Block $p(D|S)$, the scale for Block $p(S|D)$ included the words "IMPOSSIBLE, P = 0" on the left, and the words "SURE, P = 1" on the right, instead of the name of the diseases. Consequently, unlike Block $p(D|S)$, in Block $p(S \mid D)$ those judgments related to patients suffering from R were independent from those related to patients suffering from C. Half of the participants solved Block $p(D|S)$ first whereas the other half solved Block $p(S|D)$ first. A 4 s message appeared between the items of one block and the others, warning participants about the fact that the direction of their judgments had to be reversed.

Results and Discussion

All the analyses performed adopted an alpha of .05. First of all, we replicated Gluck and Bower's (1988a) apparent base-rate neglect. Specifically, participants' probability judgments about the probability of R in patients with S_1 (median = 0.63) were higher than the empirical probabilities (median = 0.49), $z = -2.74$.

Participants' judgments for Disease R given each symptom pattern (except the no symptoms one) are shown in Table A1 of the Appendix [probability judgments for C were just the complementary probabilities because of the scale used to make judgments in the $p(D \mid S)$ block]. The fact that almost all ratings were well below .5 clearly shows that participants' judgments were highly sensitive to base rates. Thus, the apparent base-rate neglect can hardly be attributed to participants' insensitivity to base rates. Participants' sensitivity to base rates with similar tasks has also been shown by Gluck and Bower (1988a) as well as by others (Estes et al. 1989; Kruschke, 1996).

The results of the conjunction fallacy tests are shown in Table 2. One of the participants told the experimenter that she had mistakenly been judging $p(S|R)$ instead of $p(S|C)$. Thus, the data of the 39 remaining participants were analyzed. A 2 (Category Frequency) × 5 (Symptom Conditional Probability) within-subjects analysis of variance (ANOVA) was performed on participants' judgments. In the case of Disease C, the levels of the latter factor were represented, ranging from 1 through 5, by S_4, S_1, S_1 and S_4; S_1 and S_3; and S_1 and S_2, respectively. In the case of Disease R, such levels were represented by the respective analogs S_1, S_4, S_4 and S_1; S_4 and S_2; and S_4 and S_3. The main effects of category frequency, $F(1, 38) = 15.93$, $MSE = 0.17$, $p < .01$; and symptom conditional probability, $F(4, 152) = 43.53$, $MSE = 0.05$, $p < .01$, were highly significant, but not the Category Frequency × Symptom Conditional Probability interaction, $F(4, 152) = 1.42$, $MSE = 0.05$, $p = .23$. Thus, as predicted, probability judgments conditional on C were systematically higher than probability judgments conditional on R along the different levels of symptom conditional probability. This somewhat counterintuitive sensitivity to category frequencies in $p(S|D)$ judgments is consistent with the claim that such judgments were biased by a process that operated in the opposite direction (i.e., from symptoms to diseases).

To test in which cases participants committed the conjunction fallacy, we conducted a series of planned contrasts between the levels of symptom conditional probability both in $p(S \mid R)$ and in $p(S|C)$ judgments. In patients suffering from Disease R, judgments for the conjunction S_4 and S_1 were significantly higher than for the constituent S_4, $F(1, 38) = 17.36$, $MSE = 0.08$, $p < .01$. However, neither the conjunction S_4 and S_2 nor the conjunction S_4 and S_3 received ratings significantly higher than S_4, $F(1, 38) = 0.03$, $MSE = 0.04$, $p = .86$; $F(1, 38) = 1.82$, $MSE = 0.09$, $p = .19$, respectively. As predicted, Symptom S_1, which is supposed to have the strongest association for R, was judged significantly more probable than the conjunction S_4 and S_1, $F(1, 38) = 22.6$, $MSE = 0.11$, $p < .01$. In patients suffering from C, only the conjunction S_1 and S_4 received ratings significantly higher than the critical constituent S_1, $F(1, 38) = 23.16$; $MSE = 0.05$, $p < .01$. Neither judgments for the conjunction S_1 and S_3 nor for the conjunction S_1 and S_2 were significantly higher than judgments for S_1, $F(1, 38) = 0.26$, $MSE = 0.13$ $p = .61$; $F(1, 38) = 0.13$, $MSE = 0.1$, $p = .72$, respectively. Also as predicted, Symptom S_4, which is supposed to have the strongest association for C, was judged significantly more probable than the conjunction S_1 and S_4, $F(1, 38) = 12.3$, $MSE = 0.07$, $p < .01$.

A remarkable result is the spectacular difference between judgments for S_1 and $S_4|C$, $p = .64$ and for S_4 and $S_1|R$, $p = .45$, on the one hand, and the corresponding empirical probabilities, $p = .1$, $p = .14$, respectively, on the other hand. These sizable overestimations make sense if we take into account participants' mean judgments for the inverse probabilities, $C|S_1$ and S_4, $p = .59$; and

Table 2
Mean Probability Judgments Corresponding to the Critical Constituents and to the Conjunctions, and Corresponding Output Activation Values for Each Item From the Neural Network Simulation in Experiment 1

	Diseases									
	C					*R*				
Judgments	S_4	S_1	S_1 and S_4	S_1 and S_3	S_1 and S_2	S_1	S_4	S_4 and S_1	S_4 and S_2	S_4 and S_3
Participants	.79	.46	.64	.49	.44	.70	.26	.45	.26	.32
Network	.74	.36	.61	.48	.39	.64	.26	.39	.24	.18

Note. C = common; R = rare; S = symptom.

S_4 and $S_1|R$, $p = .41$. Thus, these results strongly support the conversion bias hypothesis as an explanation of the conjunction fallacies. As will be shown in the next section, the referred overestimations are predicted on the basis of an associative process that always goes from symptoms (cues) to diseases (outcomes).

Simulation

We ran a very simple simulation with a version of Gluck and Bower's (1988a) model. It consisted of a bilayered neural network with four binary input nodes (S_1, S_2, S_3, and S_4) and two sigmoidal output nodes (R and C). Though the use of two sigmoidal nodes is equivalent to only one in situations of complementary probabilities, we preferred the two nodes option to avoid changes between different simulations. On the other hand, we preferred the use of sigmoidal over more common linear nodes to simplify the translation of node activities into probability judgments. In any case, both kinds of nodes produce very similar simulation results. The sigmoidal activation function used was as follows:

$$a_i = \frac{1}{1 + e^{c(-net_i)}}$$

where net_i is the net input to output node i, and c is a free parameter that determines the slope of the sigmoid-shape function. The weights of the connections from the input to the output nodes were updated according to the delta-rule. The network was trained with the same procedure and the same distribution of probabilities used to train participants. Training trials were distributed in successive blocks of 160 trials. Each block was equivalent to a single participant's training phase. The learning rate parameter was set at 5×10^{-5}. This extremely low value was adopted to provide weights very close to asymptote with sufficient training. Thus, the network was trained through an extensive number of blocks until it reached a small area of the weight space in which the weights only experienced minimal changes around a middle point.

The weights obtained at asymptote were, approximately, as follows: For node R (ordered from S_1 through S_4 input units), the vector weight was 1.17, -0.23, -1.03, -2.11, whereas, for node C, it was -1.17, 0.23, 1.03, 2.11. The free parameter c of the activation function was set to 0.49 to obtain the best fit defined as the minimum of the sum of squared errors ($SSE = 0.04$), though variations in this parameter do not qualitatively alter the simulation results. The probability judgments obtained from the neural network simulation are shown in Table A1 of the Appendix (the *SSE* measure was obtained from the difference between the ratings shown in Table A1). As can be seen, the neural network accurately predicts the apparent base-rate neglect, and, at the same time, shows the same sensitivity to base rates as participants.

The simulation results are highly consistent with the results of the conjunction fallacy tests. Table 2 shows the activation values of the output nodes for the critical items of the conjunction fallacy tests. According to such values, the conjunction S_1 and S_4 in the context of Disease R should have received higher ratings than Symptom S_4. In the context of Disease C, the conjunction S_1 and S_4 should have received higher ratings than Symptom S_1. In fact, participants committed both conjunction fallacies. Though not so confidently expected, judgments for the conjunction S_1 and S_3 should have also been higher than for S_1. This is the only prediction the data did not support. The model also correctly predicted higher judgments for $p(S_1|R)$ than for $p(S_4$ and $S_1|R)$, and higher judgments for $p(S_4|C)$ than for $p(S_1$ and $S_4|C)$ (see the corresponding output values in Table 2).

In addition, the model, together with the learning-order assumption of AL, correctly predicted systematically higher judgments for S_4, S_1, S_1 and S_4; S_1 and S_3; and S_1 and S_2 conditional on C than for their respective analogs conditional on R. It is particularly interesting to note that participants' spectacular overestimates in the case of $p(S_4$ and $S_1|R)$ and $p(S_1$ and $S_4|C)$ are highly consistent with the activation values shown in Table 2. This result strongly supports the hypothesis that participants' $p(S|D)$ judgments were strongly influenced by a process that conditionalized on symptoms rather than on diseases.

Experiment 2

According to the explanation provided for the results of Experiment 1, the conjunction fallacies obtained involved the conversion bias. That is, when participants had to make $p(S \mid D)$ judgments they tended to conditionalize on symptoms rather than on diseases because of the symptom–disease learning order used along the training phase. Since the conversion bias has played such an important explanatory role in Experiment 1, Experiment 2 was conducted to obtain stronger and independent evidence on such bias using the same learning task. This way, our explanation of the conjunction fallacies obtained in Experiment 1 would gain greater support.

In Experiment 2, we induced the conversion bias by requiring conditional probability judgments in the opposite direction to the learning order, i.e., by requiring $p(S|D)$ judgments. We also manipulated the direction of judgments during the test phase to learn the extent to which both $p(S|D)$ and $p(D|S)$ judgments are determined by the same process. More evidence in support of the conversion bias is the sensitivity to category base rates in $p(S|D)$ judgments. As can be seen in Table 1, the probabilities of S_1 through S_4 conditional on R are the same as the probabilities of S_4 through S_1 conditional on C. However, if participants commit the conversion bias, judgments in the context of the more frequent disease should be systematically higher than judgments in the less frequent one.

Although participants made conditional probability judgments in both directions in Experiment 1, they did it in very different conditions. The symptom patterns involved in $p(S|D)$ judgments were not the same as those involved in $p(D|S)$ judgments and the procedure used to request judgments in both directions was quite different. For $p(D|S)$ items, we used a scale consisting of a horizontal bar with one of the two diseases at each end. In those items, participants estimated the probability of both diseases at the same time. On the contrary, $p(S|D)$ judgments were made separately for each disease. In Experiment 2, we have overcome those contaminating aspects of the design.

In summary, the specific objectives of Experiment 2 were as follows: (a) to assess the effect of changing the direction of judgments on participants' responses, and (b) to assess the differential effect of category base rate in $p(S|D)$ and $p(D|S)$ judgments. Accordingly, we manipulated three within-subjects factors: inversion [$p(D|S)$ judgments vs. $p(S|D)$ judgments], category frequency (C vs. R) and conditional probability (.6, .4, .3 and .2). The combination of the two types of judgments with the two diseases and the four symptoms gave a total of 16 experimental items (see Table 3).

As mentioned in the introduction, Braine et al. (1990) obtained the conversion bias in Bayesian tasks. Such conversion bias was shown to be the cause of the base-rate neglect found in the inverse probability ratings.

Our Experiment 2 differs from those performed by Braine et al. (1990) in two important respects. First, those authors employed a verbal format, instead of an experiential-learning task, to provide participants with information about conditional probabilities and base rates. Thus, this procedure is not appropriate to show the capacity of experiential-learning tasks to induce the conversion bias. Second, the conversion bias we are inducing causes a sensitivity to base rates in $p(S|D)$ rather than a base-rate neglect. This is a very interesting result because it is unusual and seems counterintuitive. Up to now, the conversion bias has always been associated with the base-rate neglect.

Method

Participants. The participants were 28 psychology undergraduates from Universidad de La Laguna who volunteered to take part in the experiment as partial fulfillment of their course requirements.

Apparatus and stimuli. The apparatus and stimuli were the same as those used in Experiment 1.

Procedure. The procedure in this experiment was the same as that used in Experiment 1 except for the probability-judgment phase. To make this second phase shorter, those items in which participants had to estimate the probability of the diseases for all possible configurations of symptoms were removed. On the other hand, the scale used to make $p(D|S)$ judgments and $p(S|D)$ judgments consisted of a horizontal line segmented and numbered, from left to right, with the figures 0, 25, 50, 75 and 100. The design is shown in Table 3.

The judgment test was administered in two blocks: a $p(D|S)$ block and a $p(S|D)$ block. Each block was, in turn, divided into two subblocks: the *R* subblock, and the *C* subblock. Thus, there were a total of 4 subblocks {2[types of judgment, $p(D|S)$ vs. $p(S|D)$] × 2 (disease, *C* vs. *R*)}. Half of the participants made $p(D|S)$ judgments first, whereas the other half made them second. Orthogonally, there were also two counterbalance groups according to the order in which they answered the subblocks. One group received the subblocks in the following order *R-C-C-R*. The order for the other group was *C-R-R-C*. Besides the 16 experimental items, 16 more were included as filler items. In the condition $p(D|S)$, the filler items informed participants about the presence of a symptom and the explicit absence of the others, whereas in the condition $p(S|D)$ participants had to estimate the probability of a given isolated symptom in a hypothetical patient. As in Experiment 1, these items were included to avoid any possible linguistic confusion regarding the experimental items. At the same time, this makes the context in which experimental items were presented more similar to that of Experiment 1, so it would be easier to draw comparisons between both experiments.

Table 3
Design of Experiment 2 and Mean Probability Judgments in Each Condition

	Diseases							
	C				*R*			
Direction	S_4	S_3	S_2	S_1	S_1	S_2	S_3	S_4
$p(S\|D)$	.92	.58	.62	.33	.77	.50	.49	.20
$p(D\|S)$	.91	.58	.57	.30	.71	.47	.49	.16

Note. *C* = common; *R* = rare; $p(S|D)$ = programmed probability of the symptom conditional on the disease; $p(D|S)$ = programmed probability of the disease given the symptom; *S* = symptom; *D* = disease.

Results and Discussion

Table 3 shows the mean judgments obtained in each condition. The data were analyzed through a 2 [Inversion: $p(D|S)$ vs. $p(S|D)$] × 2 (Category Frequency: *C* vs. *R*) × 4 (Conditional Probability: .6, .4, .3, .2) within-subjects ANOVA. The main effects of category frequency, $F(1, 27) = 21.15$, $MSE = 0.08$, $p < .01$, and conditional probability, $F(3, 81) = 57.63$, $MSE = 0.11$, $p < .01$, were significant. However, the main effect of inversion was only marginally significant, $F(1, 27) = 3.19$, $MSE = 0.02$, $p = .08$. Table 3 reveals a trend, according to which $p(S|D)$ ratings tended to be slightly higher than $p(D|S)$ ratings. Though this trend is consistent with the empirical probabilities in the case of Disease *R* judgments, it is not consistent in the case of Disease *C* judgments. Thus, the marginally significant effect hardly can be attributed to a coherent discrimination between $p(S|D)$ and $p(D|S)$.

Another interesting result is the absence of interaction between category frequency and inversion, $F(1, 27) = 0.28$, $MSE = 0.02$, $p = .60$. If, as previously argued, the mechanism responsible for $p(D|S)$ judgments is also responsible for $p(S|D)$ judgments, category frequency should be expected to also affect the latter. The absence of a Category Frequency × Inversion interaction means that the effect of category frequency in $p(S|D)$ judgments was not different from the effect of category frequency in $p(D|S)$ judgments. In fact, this interpretation is confirmed by the analysis of the simple effects of category frequency in each level of inversion. The effect of category frequency was significant in condition $p(D|S)$, $F(1, 27) = 17.48$, $MSE = 0.06$, $p < .01$, as well as in condition $p(S|D)$, $F(1, 27) = 16.79$, $MSE = 0.05$, $p < .01$. Finally, none of the other possible interactions were significant.

To conclude, the conversion bias is supported by the absence of significant differences between $p(D|S)$ and $p(S|D)$ judgments, by the marked effect of category frequency in both types of judgment, and by the absence of any interaction between category frequency and inversion. Thus, the results of Experiment 2 strongly support the explanation given for the conjunction fallacies obtained in Experiment 1. At the same time, we have shown that the conversion bias can give rise to a bias opposite to the base-rate neglect, i.e., an effect of base rate on probability judgments conditional on the category [$p(S|D)$]. Such results strongly support the AL because it is rather counterintuitive and unusual.

These results cannot be easily explained by appealing to simple linguistic confusion between the probability of the diseases conditional on the symptoms, and the probability of the symptoms conditional on the diseases. The differences between such conditional probabilities were explained through clear examples in the instructions, which were not resumed until all participants had understood the explanations given. In addition, as the inversion factor received a within-subjects manipulation, participants were encouraged not to confound both types of judgment. Finally, $p(S|D)$ and $p(D|S)$ items were administered in separated blocks that were clearly announced (see Experiment 1 instructions).

Experiment 3

As we have shown through Experiments 1 and 2, the conjunction fallacy can be obtained as a by-product of the conversion bias. According to AL, participants committed the conjunction fallacy because they had to make conditional probability judgments in the opposite direction to the acquired associative links. Though Tversky and Kahneman (1983, p. 303) acknowledged that this sort of conjunction fallacy could happen in some problems, they nonetheless considered that it was clearly implausible for others. In general, they conceived that, in *Linda*-like problems, associations are likely to be directed in the same direction as the probability judgment. For example, in the *Linda* problem, there would be a high excitatory association from Linda's personality traits to the feminist statement and an inhibitory association from Linda's personality traits to the bank teller statement. The same occurs in other problems such as the *Bill* or the *Wimbledon* problems (Tversky & Kahneman, 1983).

It would be very interesting if experiential-learning tasks, such as Gluck and Bower's (1988a), could be adapted to induce the conjunction fallacy in probability judgments consistent with the direction of the acquired associative links. Thus, the conjunction fallacy obtained would be more readily comparable to the more common *Linda*-like problems. As stated in the introduction, an important advantage of this is that associative models, as well as other quantitative accounts, could be tested to a greater extent, and, thus, the precise conditions in which biases occur could be further specified.

The aim of Experiment 3 was to explore whether the conjunction fallacy could be induced in probability judgments consistent with the learning order used along the training phase. As in the previous experiments, participants had to learn to diagnose two diseases, *A* and *B*, from four symptoms, S_1 through S_4. The probabilities of the symptoms conditional on the diseases were changed to reach more extreme values of diagnostic information (see Table 4). Unlike Experiments 1 and 2, Diseases *A* and *B* were equiprobable ($p = .5$), and they were not mutually exclusive. Thus, there were patients who simultaneously suffered from both diseases. The probability of Disease *A* and of Disease *B* were independent from each other. Thus, the probability of the conjunction *A* and *B* was equal to the product of the probabilities of each disease ($p = .5 \times .5 = .25$). However, as patients with neither disease were excluded from the training trials, the empirical probabilities came to be, with slight deviations, as follows: $p(A) = p(B) = .66$, $p(A \text{ and } B) = .33$. In patients suffering from both *A* and *B*, the symptoms were independently produced by each disease (i.e., there was no interaction). Consequently, the probabilities of S_1 through S_4 conditional on the conjunction of diseases *A* and *B* were as shown in Table 4.

To assess whether or not participants committed the conjunction fallacy, they were required to judge the probability of items such as $p(B|\text{only } S_1)$ and $p(A \text{ and } B|\text{only } S_1)$ in the test phase. According to the strategy used in more traditional demonstrations of the conjunction fallacy, the greater the difference between the $S_x \rightarrow A$ and the $S_x \rightarrow B$ associative strengths, the greater the chances of the bias to occur. As can be easily inferred from Table 4, after the training phase, Symptom S_1 should be strongly associated with Disease *A* and inhibitorily, or very weakly, associated with Disease *B*. Thus, judgments for the conjunction *A* and *B* should be higher than for the constituent *B*.

In Experiment 1, the conjunction fallacies were directly predicted from the output node activities of the neural network because, according to AL, $p(S|D)$ judgments should be strongly influenced by associative links consistent with the learning order. Since in Experiment 3 the direction of judgments is consistent with the direction of the presumed associative links, the conjunction fallacies cannot be explained as a by-product of a conversion bias. However, we will show through a simulation that the conjunction fallacies obtained in Experiment 3 can be explained by postulating an integration process that takes the two output values (one for each disease) of the associative-learning mechanism and computes a sort of average response to make conjunction judgments. This is still compatible with AL, for one of its assumptions is that the output of the associative-learning mechanism can also feed another nonextensional process, and, thus, can indirectly produce the bias.

There is a large body of research on judgmental integration supporting the averaging-process hypothesis for combining the probability of outcomes (see, e.g., Birnbaum & Mellers, 1983; or, more related to the conjunction fallacy, Carlson & Yates, 1989; Fantino, Kulik, Stolarz-Fantino, & Wright, 1997; Hampton, 1997; Yates & Carlson, 1986). The details of the averaging model used

Table 4
Programmed Probabilities, Observed Judgments, Simulated Judgments Obtained with the Double-Component Model, the Double-Component Model, the MINERVA-DM, and the Corresponding Empirical Probabilities of Experiment 3*

	Diseases														
	A					*A* and *B*					*B*				
Source	S_1	S_2	S_3	S_4	All	S_1	S_2	S_3	S_4	All	S_1	S_2	S_3	S_4	All
$p(S\|D)$	.80	.50	.30	.20	.02	.84	.65	.65	.84	.30	.20	.30	.50	.80	.02
OJ	.74	.70	.42	.33	.71	.43	.41	.43	.44	.73	.34	.45	.65	.83	.71
DCM	.81	.64	.47	.33	.77	.43	.50	.50	.43	.77	.33	.47	.64	.81	.77
DCM*	—	—	—	—	—	.57	.55	.55	.57	.77	—	—	—	—	—
MDM	.85	.71	.43	.32	.83	.27	.23	.23	.27	.69	.32	.41	.68	.85	.81
EP	.94	.74	.35	.11	.94	.06	.06	.11	.06	.86	.11	.31	.76	.94	.92

Note. S = symptom; D = disease; $p(S|D)$ = programmed probabilities; OJ = observed judgments; DCM = double-component model; DCM* = double-component model ($\alpha = .5$); MDM = MINERVA-decision making; EP = empirical probabilities.

in our simulation will be described in the *Simulation* section below.

Method

Participants. The participants were 37 psychology undergraduates from Universidad de Málaga who volunteered to take part in the experiment as partial fulfillment of their course requirements.

Apparatus and stimuli. Apparatus and stimuli were the same as those used in the previous experiments.

Procedure. As the procedure used in this experiment was the same as in Experiments 1 and 2, we will only describe the different aspects. We removed from the instructions any reference to judgments about probabilities of the symptoms conditional on the diseases because the test phase only consisted of $p(D|S)$ items. Given participants' remarkable trend to conditionalize on symptoms in Experiments 1 and 2, we no longer considered the inclusion of $p(S \mid D)$ filler items necessary. The instructions given in the previous experiments to discriminate between the explicit absence of symptoms, and the absence of information about the presence or absence of such symptoms, were now referred to the diseases in Experiment 3. In the present experiment, such instructions were intended to prevent participants from confusing the probability of a given disease [e.g., $p(B|\text{only } S_1)$] with the probability of such disease in the absence of the other [i.e., $p(B \text{ and no } A|\text{only } S_1)$]. Otherwise, the conjunction fallacy could be no more than a mere misinterpretation of the items. We also warned participants that the conjunction items referred to suffering simultaneously from the two diseases rather than to suffering from one or the other.

Regarding the training phase, we altered the programmed probabilities, the diagnostic responses that could be performed, and the number of training trials. Table 4 shows the programmed probabilities. To diagnose Disease *A*, or *B*, participants had to press either Key 1, or Key 2, respectively. To diagnose the conjunction of both diseases, participants had to press Key 1 and Key 2 successively in any order. After the diagnosis, the enter key had to be pressed. The number of training trials was increased up to 180 because the learning task was presumably harder than before. Note that, in Experiment 3, participants had to learn to infer three possible outcomes (*A*, *B*, and *A* and *B*) on each training trial.

In every item of the test phase, participants had to judge the probability of the diseases conditional on the symptoms [$p(D|S)$]. The test phase consisted of three different types of items distributed in three different blocks. In two of the three blocks, participants had to judge either the probability of Disease *A* [$p(A|S_x)$] or the probability of Disease *B* [$p(B|S_x)$]. Thus, these two blocks included constituent items. In the third block, participants had to judge the probability of the conjunction of both diseases [$p(A \,\&\, B|S_x)$]; it was, thus, the conjunction block. There were five items in each block: one for each symptom plus another one for the conjunction of the four symptoms [e.g., $p(A \,\&\, B|S_1 \text{ and } S_2 \text{ and } S_3 \text{ and } S_4)$]. Except for the all-symptoms item, the remaining four consisted of probabilities conditional on the presence of one of the symptoms and the explicit absence of the remaining three. Thus, there were a total of 15 experimental items: 3 Item Types × 5 Symptom Patterns. We also included 10 filler items divided into two blocks. In one block, judgments were required about the probability of Disease *A* in the absence of Disease *B* conditional on the same symptom patterns as in the experimental items [$p(A \text{ and no } B|S_x)$]. In the other block, participants had to judge, instead, the probability of Disease *B* in the absence of Disease *A* conditional, again, on the same symptom patterns [$p(B \,\&\, \text{no } A|S_x)$]. These filler items were included to encourage participants to discriminate the probability of a given disease from the probability of such disease in the absence of the other. Half of the participants answered the filler items before the experimental ones, whereas the other half proceeded in the opposite order. Orthogonally, half of the participants answered the conjunction items first, whereas the other half answered the constituent items first. Each block was preceded by a message that specified which type of items came next. Such message was displayed in the center of the PC screen for 4 s.

We used the same visual cues as in Experiments 1 and 2 to promote participants' discrimination between the absence of a disease and the absence of information about the presence or absence of such disease.

Results and Discussion

Table 4 shows participants' mean judgments for each experimental item. For each symptom pattern (a total of five), participants judged the probability of *A*, of *B*, and of the conjunction *A* and *B*. Consequently, we have tested the conjunction fallacy in five different conditions. However, such conditions are duplicated. Consider, for example, the comparison between judgments in the $p(A|\text{only } S_4)$ item and judgments in the $p(A \text{ and } B|\text{only } S_4)$ item to test whether the conjunction fallacy takes place. If we take into account that items such as $p(A \mid \text{only } S_4)$ and $p(A \text{ and } B|\text{only } S_4)$ implement the same conditions as items $p(B|\text{only } S_1)$ and $p(A \text{ and } B|\text{only } S_1)$, respectively (see Table 4), we reach the conclusion that the same conjunction fallacy has been tested twice. This is advantageous because we can work with more reliable measures by averaging between the two judgments of the same condition. Thus, the statistical analyses have been applied to the average judgments obtained from the two items of the same condition. However, for simplicity, we will only use Disease *A* as a notational system to refer to the conditions under which judgments were made. For example, judgments in the $p(A|\text{only } S_4)$ condition refer to the average between judgments in the $p(A \mid \text{only } S_4)$ item and judgments in the $p(B|\text{only } S_1)$ item. Analogously, judgments in the $p(A \text{ and } B|\text{only } S_4)$ condition refer to the average between judgments in the $p(A \text{ and } B|\text{only } S_4)$ item and judgments in the $p(A \text{ and } B|\text{only } S_1)$ item. Thus, the comparison between $p(A \mid \text{only } S_4)$ and $p(A \text{ and } B|\text{only } S_4)$ conditions refers to the contrast of the average between judgments in $p(A|\text{only } S_4)$ and $p(B|\text{only } S_1)$ items against the average between judgments in $p(A \text{ and } B|\text{only } S_4)$ and $p(A \text{ and } B|\text{only } S_1)$ items. The same logic applies to the remaining conditions except for the $p(A \text{ and } B|S_1 \text{ and } S_2 \text{ and } S_3 \text{ and } S_4)$ condition because it only included a single item.

A 2 (Disease: single vs. conjunction) × 5 (Symptom Configuration: 1 through 5) within-subjects ANOVA was performed on participants' probability judgments. The disease factor allowed us to compare the probability judgments in the $p(A|S_x)$ condition with probability judgments in the $p(A \text{ and } B|S_x)$ condition. The effects of disease, $F(1, 36) = 29.7$, $MSE = 0.03$, $p < .01$, and of symptom configuration, $F(4, 144) = 28.96$, $MSE = 0.05$ $p < .01$, were significant. The Disease × Symptom Configuration interaction, $F(4, 144) = 16.99$, $MSE = 0.04$ $p < .01$, was also significant. This interaction is presumably related to the fact that the difference between the single and the conjunction judgments do not have the same sign in all symptom configuration levels. This is hardly surprising if we bear in mind the logic with which the conjunction tests were constructed. For example, in the case of conditions $p(A|\text{only } S_1)$ and $p(A \text{ and } B|\text{only } S_1)$, we are comparing probability judgments for a very strongly associated disease with probability judgments for the conjunction of such disease with a negatively associated one. Thus, judgments for the former are expected to be higher than for the latter. Conversely, in the case of $p(A|\text{only } S_4)$ and $p(A \text{ and } B|\text{only } S_4)$, we are comparing probability judgments for a negatively associated disease with probability judgments for the conjunction of such disease with a very strong and positively associated disease. Thus, judgments for the former are expected to be lower than for the latter.

To confirm these predictions, we analyzed the single effect of disease in each level of symptom configuration. As expected, participants' judgments for $p(A|\text{only } S_1)$ were significantly higher than for $p(A \text{ and } B|\text{only } S_1)$, $F(1, 36) = 55.57$, $MSE = 0.04$, $p <$ .01. This result parallels those obtained by Tversky and Kahneman (1983) with more traditional tasks. Also as expected, participants' judgments for $p(A \text{ and } B|\text{only } S_4)$, were significantly higher than for $p(A|\text{only } S_4)$, $F(1, 36) = 5.14$, $MSE = 0.03$, $p = .03$. The latter result implies the commission of a conjunction fallacy. This conjunction fallacy was the most confidently expected.

On the other hand, we did not obtain the conjunction fallacy in probability judgments conditional on only Symptom S_3: $p(A \text{ and } B|\text{only } S_3) = .42$; $p(A|\text{only } S_3) = .43$; $F(1, 36) = 0.08$, $MSE = 0.04$, $p = .77$. The reason for this result may be that the difference between the $S_3 \rightarrow A$ associative strength and the $S_3 \rightarrow B$ associative strength is not great enough to induce the conjunction fallacy. However, this interpretation may be viewed, at first glance, as inconsistent with the fact that judgments in the $p(A|\text{only } S_2)$ condition ($M = .68$) were significantly higher than judgments in the $p(A \text{ and } B|\text{only } S_2)$ condition ($M = .42$), $F(1, 36) = 30.25$, $MSE = 0.04$, $p = < .01$. Note that, according to the programmed probabilities, the difference between the $S_3 \rightarrow A$ and the $S_3 \rightarrow B$ associative strengths should be the same as that between the $S_2 \rightarrow A$ and the $S_2 \rightarrow B$ associative strengths. Thus, it remains to be explained why $p(A \mid \text{only } S_3)$ and $p(A \text{ and } B|\text{only } S_3)$ judgments did not differ, and, at the same time why $p(A \mid \text{only } S_2)$ judgments were higher than $p(A \text{ and } B|\text{only } S_2)$ judgments.

As stated in the introduction of Experiment 3, the conjunction judgments could be based on an averaging process that takes the independent judgments for Diseases *A* and *B* as input. However, in such a process, the least likely constituent could receive more weight than the most probable one. This differential weighting could satisfy the explanation demanded above. Consistent with this assumption, the results in Table 4 show that conjunction judgments are always much closer to judgments for the less likely constituent than to judgments for the more likely constituent. For example, judgments for items $p(A|\text{only } S_4)$, $p(B|\text{only } S_4)$, and $p(A \text{ and } B|\text{only } S_4)$ were .33, .83, and .44, respectively. On the other hand, judgments for items $p(A \mid \text{only } S_1)$, $p(B|\text{only } S_1)$, and $p(A \text{ and } B|\text{only } S_1)$ were .74, .34, and .43, respectively. Interestingly, and also according to the averaging hypothesis, conjunction judgments were never lower than judgments for the less likely constituent.

Another interesting result is that judgments in conditions $p(A|S_1 \text{ and } S_2 \text{ and } S_3 \text{ and } S_4)$ ($M = .71$) and $p(A \text{ and } B|S_1 \text{ and } S_2 \text{ and } S_3 \text{ and } S_4)$ ($M = .73$) were nearly identical, $F(1, 36) = 0.21$, $MSE = 0.04$, $p = .65$. This result also confirms the averaging hypothesis because judgments for items $p(A|S_1 \text{ and } S_2 \text{ and } S_3 \text{ and } S_4)$ and $p(B|S_1 \text{ and } S_2 \text{ and } S_3 \text{ and } S_4)$ were also nearly identical. At the same time, we can infer that participants did not conflate the absence of information about one of the diseases with the absence of such disease since judgments in the $p(A|S_1 \text{ and } S_2 \text{ and } S_3 \text{ and } S_4)$ condition were much higher than the empirical probabilities $p(A \text{ and no } B|S_1 \text{ and } S_2 \text{ and } S_3 \text{ and } S_4) = .08$ and $p(B \text{ and no } A|S_1 \text{ and } S_2 \text{ and } S_3 \text{ and } S_4) = .06$. Moreover, participants seemed to find it very difficult to judge the probability of one disease in the absence of the other. This claim is based on participants sizable overestimation of the filler item $p(A \text{ and no } B|S_1 \text{ and } S_2 \text{ and } S_3 \text{ and } S_4) = .66$. In summary, these results allow us to rule out the possibility that the conjunction fallacies found could be attributed to interpreting $p(A|S_x)$ as $p(A \text{ and no } B|S_x)$.

Finally, participants' conjunction judgments in the test phase were remarkably at variance with the corresponding empirical probabilities. This is also consistent with the averaging process hypothesis. Specifically, all conjunction judgments conditional on isolated symptoms were above .40, whereas all the corresponding empirical probabilities were below .12 (see Table 4). Also, there were other deviations concerning some judgments about the probability of single diseases. However, such deviations seemed to be of a quite different nature. First, they nicely conformed to the well-known phenomenon of conservatism: probability judgments tended to be less extreme than empirical probabilities. Second, they were much smaller than the deviations of conjunction judgments.

In summary, Experiment 3 shows that the conjunction fallacy can be induced by experiential-learning tasks in which the learning order is consistent with the direction of conditional probability judgments in the test phase. We have also shown that the greater the difference between the summed associative strength of the present cues for one constituent and the summed associative strength of the same cues for the other constituent, the greater the chances of obtaining a conjunction fallacy. In such cases, judgments for the conjunction are higher than for the constituent with the weaker association.

Simulation

The aim of the simulation is to show that the same neural network used in Experiment 1, plus an averaging model, can explain the conjunction fallacy and can account for the overall pattern of results from Experiment 3. According to our proposal, single-disease judgments are based on the activation value of the output node representing the disease. Alternatively, the conjunction judgments are based on some averaging measure of the two output nodes. The averaging measure consists of a simple weighted sum of the two output values in which the lower output value is affected by a greater weight than the higher output value.

The training procedure was the same as in Experiment 1. Again, a very small learning rate was used ($lr = 5 \times 10^{-5}$).

Preasymptotic weights were used to simulate participants' judgments because they provided a better quantitative fit than asymptotic weights. The only relevant effect of choosing preasymptotic weights is that probability judgments conditional on S_1 and probability judgments conditional on S_4 get slightly closer to probability judgments conditional on S_2 and probability judgments conditional on S_3, respectively. The assumption that participants did not reach asymptote is more than reasonable given the difficulty of the learning task. It is also possible that an annealing process prevents participants from converging to the best solution (Kruschke & Johansen, 1999). It is worthy to note, however, that using preasymptotic rather than asymptotic weights only improves the simulation of participants' judgments at a quantitative level. The presence and absence of the conjunction fallacy are well predicted with both sets of weights.

As the weight vector obtained for nodes *A* and *B* are symmetrical, we only report the former. At asymptote, the weights obtained for Symptoms S_1 through S_4 were, approximately, 2.87, 1.08, -0.09, and -1.66, respectively. The corresponding preasymptotic weights used to simulate probability judgments were 2.13, 0.83, -0.16, and -1.02, respectively.

According to the averaging process, the conjunction judgments were computed through the following equation:

$$p(A \text{ and } B) = G[p(A), p(B)] \times (1 - \alpha) + S[p(A), p(B)] \times \alpha,$$

where G is a function whose output is the greatest of two probability judgments [$p(A)$ and $p(B)$]; S is a function whose output is the smallest of two probability judgments; and alpha is a weighting factor such that $0 \leq \alpha \leq 1$. Thus, there were three free parameters: c, α, and the amount of training. The best quantitative fits were obtained with $c = 0.69$, and $\alpha = .8046$ ($SSE = 0.03$). The exact amount of training blocks is not informative because it depends on the size of lr, which was set at a very low value to obtain representative or central vector weights at any moment. As stated in the *Simulation* section of Experiment 1, the qualitative pattern of results is largely independent of the value adopted for c. Specifically, the predictions concerning where the conjunction fallacies should be met and where they should not are not altered. The simulated judgments are shown in Table 4.

The simulation data reproduced the pattern of differences between probability judgments obtained across the different conjunction fallacy tests. For example, the model correctly predicted participants' conjunction fallacies, since it produced a higher rating in the $p(A \text{ and } B|\text{only } S_4)$ condition, $p = .43$, than in the $p(A|\text{only } S_4)$ condition, $p = .33$. The model also predicted judgments much higher in the $p(A|\text{only } S_1)$ condition than in the $p(A \,\&\, B \mid \text{only } S_1)$ condition (see Table 4), which is consistent with the observed judgments. Not surprisingly, the simulation correctly reproduced the absence of differences between judgments in the $p(A|S_1 \text{ and } S_2 \text{ and } S_3 \text{ and } S_4)$ condition and judgments in the $p(A \text{ and } B|S_1 \text{ and } S_2 \text{ and } S_3 \text{ and } S_4)$ condition.

It is interesting to note that, though the difference between the $S_3 \rightarrow A$ and $S_3 \rightarrow B$ associative strengths is the same as the difference between the $S_2 \rightarrow A$ and $S_2 \rightarrow B$ associative strengths, the simulated judgments in the $p(A|\text{only } S_3)$ and the $p(A \text{ and } B|\text{only } S_3)$ conditions were nearly identical, whereas the simulated judgment in the $p(A|\text{only } S_2)$ condition was substantially higher than in the $p(A \,\&\, B|\text{only } S_2)$ condition. Of course, this latter result is highly dependent on the value assigned to alpha. However, the following aspects of the simulation data are highly stable along variations of alpha: (a) the conjunction is never judged to be less probable than the less likely constituent; (b) judgments in the $p(A|S_1 \text{ and } S_2 \text{ and } S_3 \text{ and } S_4)$ and in the $p(A \text{ and } B|S_1 \text{ and } S_2 \text{ and } S_3 \text{ and } S_4)$ conditions are equal for any alpha; (c) the difference between judgments in the $p(A \text{ and } B|\text{only } S_4)$ and in the $p(A|\text{only } S_4)$ conditions is higher than that between judgments in the $p(A \text{ and } B|\text{only } S_3)$ and in the $p(A|\text{only } S_3)$ conditions; (d) and, symmetrically, the difference between judgments in the $p(A \text{ and } B|\text{only } S_1)$ and in the $p(A|\text{only } S_1)$ conditions is higher than that between judgments in the $p(A \text{ and } B|\text{only } S_2)$ and in the $p(A|\text{only } S_2)$ conditions; (e) for almost any alpha, judgments in the $p(A \text{ and } B|S_1 \text{ and } S_2 \text{ and } S_3 \text{ and } S_4)$ condition should be higher than all the remaining conjunction judgments. To help appreciate all these assertions, Table 4 includes the simulation data that should be obtained for $\alpha = 5$. This pattern of results is directly related to the averaging component of the model and mirrors the pattern of observed judgments. Therefore, participants' judgments strongly support the averaging-process hypothesis postulated in the model.

Thus, we have shown that the double-component model does an impressive job in reproducing the qualitative pattern of observed judgments, in that it correctly predicts the cases in which the conjunction fallacy will be found, and in that it quantitatively fits such judgments.

General Discussion

Summary

In Experiments 1 through 3 we have provided empirical support for the AL for biases in probability judgment by using experiential-learning tasks to induce some biases in the probability judgments lately required.

In Experiment 1, we used Gluck and Bower's (1988a) diagnostic-learning task to replicate their apparent base-rate neglect and to induce some conjunction fallacies as a by-product of the conversion bias. The conversion bias was induced by asking participants to make probability judgments in the opposite direction to the learning order. In accordance with one assumption of AL, the symptom–disease learning order used in the training phase should have primed the use of symptom–disease associations during the judgment phase. Thus, participants should be strongly inclined to make judgments in the symptom–disease direction even if they have to make $p(S|D)$ judgments. A simulation with a simple neural network similar to Gluck and Bower's allowed us to make accurate predictions about the specific cases in which a conjunction fallacy should and should not be expected.

In Experiment 2, we looked for stronger and independent evidence of the conversion bias using the same diagnostic-learning task as in Experiment 1. We found that participants committed the conversion bias by showing that judgments were unaffected by the inversion of the conditional probability judgments, and that both $p(D|S)$ and $p(S|D)$ judgments were equally affected by the frequency of D. Thus, the results of Experiment 2 strongly support the explanation given for the conjunction fallacies of Experiment 1.

Experiment 3 shows that the conversion bias is not a necessary condition to obtain the conjunction fallacy. In such an experiment, we induced the conjunction fallacy in conditional probability judgments in which the direction was consistent with the cue–outcome learning order of the training phase. We showed that the greater the difference between the summed associative strength of the present cues for one disease and the summed associative strength of the same cues for the other disease, the greater the chances of obtaining a conjunction fallacy. In the more extreme cases, participants judged the conjunction more probable than the least associated disease. In addition, we showed that a double-component model consisting of the neural network used in Experiment 1, and a very simple averaging model that integrates the network output activities in a single response, could accurately account for the overall pattern of results, and, importantly, correctly predict participants' conjunction fallacies.

Alternative Accounts of the Results

Recently, there has been an increase in the amount of quantitative models designed to explain biases in probability judgment. For example, as stated in the introduction, the PMM theory, the BIAS model, and the MDM model have caused an interesting impact in the field. It remains to be discussed, then, whether the double-component model is only one more in the growing list or, alternatively, possesses some remarkable advantages over the others. We will briefly examine this issue by considering the competence of the models to account for the data of our experiments.

PMM has been proposed to explain how people choose an outcome between two alternatives that have an uncertain status. According to PMM, choosing between two alternatives is the result of a comparison process based on the presence and/or absence of a series of cues that have a specific validity to predict the uncertain status. However, PMM is silent regarding how knowledge about cue predictive value is acquired. Thus, the application of PMM to our data is very limited.

BIAS explains many biases as a consequence of an aggregation phenomenon that takes place as a result of memory processes. However, BIAS is neutral regarding whether aggregation takes place at the moment of storage or at the moment of retrieval, and is not committed to a specific aggregation procedure. The neural network used in our simulations is an aggregation model in which the aggregation process takes place at the moment of storage. Thus, this model can be viewed as a functional implementation of BIAS (see Fiedler, 1996, for further information about the relationship between BIAS and neural networks that use the delta rule). The specific aggregation procedure used by Fiedler is computationally equivalent to a Hebbian bilayered neural network with linear output units and with normalized weights. Such a neural network would predict the apparent base-rate neglect found in Experiment 1 and replicated in Experiment 2 (see Table 3) but, contrary to participants' overall judgments, it would be largely insensitive to base rates. Therefore, BIAS's coherence with the data of Experiments 1 and 2 depends crucially on the aggregation procedure used.

Regarding MDM, it is an adaptation of Hintzman's (1984, 1988) MINERVA 2 exemplar memory model to calculate conditional probabilities of the form $p(H_i|D_j)$. As stated in the introduction, MDM assumes a certain commitment with extensional logic; i.e., probabilities are assigned on the basis of the relationship between the extensions of sets. The extensions of sets are calculated by means of a mechanism that works basically as an event counter.

The simulations we have performed with MDM as well as the analysis of its basic processing principles revealed that the double-component model is largely superior regarding its competence to account for our data. This superiority is manifested in, at least, three aspects: (a) MDM does not explain the apparent base-rate neglect obtained in Experiments 1 and 2; (b) the quantitative fit provided by MDM with respect to participants' $p(D|S)$ judgments in Experiments 1 and 3 is remarkably poorer than that provided by the double-component model; (c) the simulated data provided by MDM for Experiment 3 is qualitatively different from the pattern observed in participants' judgments. The details of the simulations are specified in the Appendix.

But, what is the reason for the superiority of the double-component model with respect to MDM? The answer to this question has two parts. On the one hand, the results obtained in Experiments 1 and 2 (the apparent base-rate neglect), together with those obtained by other authors, point to the selective-learning mechanism of the neural network as a crucial aspect (see also Estes et al., 1989, for further simulation work supporting this claim). According to the selective-learning mechanism, the weight received by a given cue or feature reflects its relative validity to predict the category rather than the objective frequency of feature–category pairings. On the contrary, MDM's computations are more related to the objective frequency of feature–category pairings. This is because all features have the same likelihood of being stored in memory and make the same contribution in the frequency count of any event. There is an overwhelming set of data showing that when participants go through an experiential-learning task, their probability judgments reflect the relative validity of cues rather than the objective conditional probabilities or frequency counts (see, e.g., Cobos et al., 2000; Cobos et al., 1993; Estes et al., 1989; Kruschke, 1996; Kruschke & Johansen, 1999; Lagnado & Shanks, 2002; Myers et al., 1994; Nosofsky et al., 1992; Shanks, 1990). Moreover, Lagnado and Shanks have recently shown that participants' sensitivity to relative validity relies on their conflating "probability" with "predictiveness." In several experiments they found that the sensitivity to relative validity tends to disappear when probability judgments are requested in a frequency format.

Perhaps MDM would improve and, thus, account for our data if it included a selective-learning mechanism allowing each feature to have a different weight in the event-frequency calculus. Of course, this has the danger of simply exchanging an exemplar-memory model for an associative-learning model.

The second part of the answer is that MDM reduces to Bayes's theorem in the limit (Dougherty et al., 1999). It is also true that MDM deviates from Bayes's theorem if certain free parameters are set so that the event-counter mechanism produces some false recognitions of a given object as well as some incorrect rejections of such an object. However, the deviations produced by the imperfect processing of the event-counter mechanism are not the kind of deviations necessary to explain both the apparent base-rate neglect obtained in our experiments and the qualitative pattern of the data obtained in Experiment 3 (see the Appendix).

Potential Extrapolation of the AL Framework

In most experiments of conjunction fallacy, conversion bias, and base-rate neglect, predictive relationships between events are framed within a causal or categorization context. The information participants have about such relationships is the basis on which probability judgments are made. The possibility of extrapolating the AL account to other tasks used to study biases depends on the origin of such information. In many of the problems, the information about the predictive relationships between the task events depends on participants' previous knowledge that has been acquired, fundamentally, in their natural environment. For example, in the *Linda* problem, participants have to make probability judgments on the basis of their previous knowledge about causal relationships between Linda's personality traits and the activities she might potentially carry out. Such knowledge might have been acquired through repeated exposure to the presence of the events involved in the *Linda* problem. As it is well known, associative-learning models have been successfully employed to understand the learning of predictive relationships (including causal learning, and category learning) in experiential-learning situations (see Shanks, 1995 for a review). Consequently, it may be argued that associative-learning mechanisms are largely involved in the previous acquisition of the knowledge on which judgments are based in *Linda*-like problems. Thus, the explanation of biases in this kind of task may well be within the scope of AL because, according to such a framework, biases take place if the contents of a given probability-judgment problem have been previously involved in an experiential-learning situation.

In other experiments, however, participants do not bring any previous relevant knowledge about the relationships between the task contents. In such situations, information is provided through verbal instructions, and using a numerical format. For instance, in the *taxicab problem* described in Footnote 1, participants receive numerical information about the number of cabs of each color that operate in a city and about the reliability of a witness who claims that a blue cab was involved in a hit-and-run accident. Given this information, participants have to judge the probability of the taxi involved in the accident being blue. Despite that the conversion bias and the base-rate neglect have been obtained with this problem, the AL account cannot be extrapolated to these sorts of situations because the information provided to participants has not been acquired through an associative-learning mechanism.

Concluding Comments

Of course, there are many important issues out of the scope of the present article that remain to be dealt with to reach a better understanding of the circumstances in which biases do and do not occur. Two general issues need to be addressed: (a) which factors modulate the intervention of learning processes other than AL in knowledge acquisition? (b) which aspects of the judgment phase modulate the intervention of extensional reasoning processes? The first question has only been approached very recently by researchers devoted to the study of predictive learning (see, e.g., Erickson & Kruschke, 1998; Price & Yates, 1995; Roberts & MacLeod, 1995; Shanks, 1995; Shanks & Darby, 1998). The second question can be viewed as part of a line of research with a larger tradition in which researchers try to delineate those factors that make people change from an intuitive reasoning strategy to an extensional reasoning one (Fiedler, 1988; Gigerenzer, 1991; Tversky & Kahneman, 1983).

Whatever the answers to these general issues will be, they will only be clarified if theories are specified in terms of models of cognitive processes. We hope to have shown throughout this paper how much our understanding of biases in probability judgment can be improved when quantitative, associative approaches and experiential-learning tasks are used.

References

Agnoli, F., & Krantz, D. H. (1989). Suppressing natural heuristics by forma instructions: The case of the conjunction fallacy. *Cognitive Psychology, 21,* 515–550.

Baker, A. G., Mercier, P., Vallée-Tourangeau, F., Frank, R., & Pan, M. (1993). Selective associations and causality judgments: The presence of a strong causal factor may reduce judgments of a weaker one. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 414–432.

Birnbaum, M. H. & Mellers, B. A. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology, 45,* 792–804.

Braine, M. D. S., Connell, J., Freitag, J., & O'Brien, D. P. (1990). Is the base rate fallacy an instance of asserting the consequent? In K. J. Gilhooly, M. T. G. Keane, R. H. Logie, & G. Erdos (Eds.), *Lines of thinking: Reflections on the psychology of thought* (Vol. 1, pp. 165–180). Chichester, England: Wiley.

Carlson, B. W. & Yates, J. F. (1989). Disjunction errors in qualitative likelihood judgment. *Organizational Behavior and Human Decision Processes, 44,* 368–379.

Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory & Cognition, 18,* 537–545.

Cobos, P. L., Caño, A., López, F. J., Luque, J. L., & Almaraz, J. (2000). Does the type of judgement required modulate cue competition? *The Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology, 53B,* 193–207.

Cobos, P. L., López, F. J., Rando, M. A., Fernández, P., & Almaraz, J. (1993). Connectionism and probability judgment: Suggestions on biases. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society* (pp. 342–346). Hillsdale, NJ: Erlbaum.

Dickinson, A., & Shanks, D. R. (1985). Animal conditioning and human causality judgment. In L. G. Nilsson, & T. Archer (Eds.), *Perspectives on learning and memory* (pp. 167–191). Hillsdale, NJ: Erlbaum.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory process model for judgments of likelihood. *Psychological Review, 106,* 180–209.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology General, 127,* 107–140.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 556–571.

Fantino, E., Kulik, J., Stolarz-Fantino, S., & Wright, W. (1997). The conjunction fallacy: A test of averaging hypotheses. *Psychonomic Bulletin and Review, 4,* 96–101.

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research, 50,* 123–129.

Fiedler, K. (1996). Explaining and simulating judgment biases as an aggregation phenomenon in probabilistic multiple-cue environments. *Psychological Review, 103,* 193–214.

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.

Gavanski, I., & Hui, C. (1992). Natural sample spaces and uncertain belief. *Journal of Personality and Social Psychology, 63,* 766–780.

Gerolin, M., & Matute, H. (1999). Bidirectional associations. *Animal Learning & Behavior, 27,* 42–49.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". In W. Stroebe & M. Hewstone (Eds.), *European Review of Social Psychology: Vol. 2* (pp. 83–115). Chichester, England: Wiley.

Gigerenzer, G. (1993). The bounded rationality of probabilistic mental models. In K. I. Manktelow & D. E. Over (Eds.), *Rationality: Psychological and philosophical perspectives* (pp. 284–313). London: Routledge.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review, 103,* 592–596.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650–669.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats: *Psychological Review, 102,* 684–704.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528.

Gluck, M. A., & Bower, G. H. (1988a). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227–247.

Gluck, M. A., & Bower, G. H. (1988b). Evaluating an adaptive network model of human learning. *Journal of Memory and Language, 27,* 166–195.

Hampton, J. A. (1997). Conceptual combination: Conjunction and negation of natural concepts. *Memory & Cognition, 25,* 888–909.

Hinton, G. E. (1990). Mapping part–whole hierarchies into connectionist networks. *Artificial Intelligence, 46,* 47–76.

Hintzman, D. L. (1984). MINERVA A2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16,* 96–101.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528–551.

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

Kahneman, D., & Tversky, A. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153–160). England: Cambridge University Press.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103,* 582–591.

Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior: Aversive stimulation* (pp. 9–33). Miami: University of Miami Press.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44.

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 3–26.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1083–1119.

Lagnado, D. A., & Shanks, D. R. (2002). Probability judgment in hierarchical learning: a conflict between predictiveness and coherence. *Cognition, 83,* 81–112.

López, F. J., Cobos, P. L., Caño, A., & Shanks D. R. (1998). The rational analysis of human causal and probability judgment. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 314–352). Oxford, England: Oxford University Press.

Matute, H., Arcediano, F., & Miller, R. (1996). Test question modulates cue competition between causes and between effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 182–196.

McMullen, M. N., Fazio, R. H., & Gavanski, I. (1997). Motivation attention and judgment: A natural sample spaces account. *Social Cognition, 15,* 77–90.

Miller, R. R., & Barnett, R. C. (1993). The role of time in elementary associations. *Current directions in Psychological Science, 2,* 106–111.

Myers, J. L., Lohmeier, J. H., & Well, A. D. (1994). Modeling probabilistic categorization data: Exemplar memory and connectionist nets. *Psychological Science, 5,* 83–89.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 211–233.

Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review, 101,* 587–607.

Politzer, G., & Noveck, I. A. (1991). Are conjunction rule violations the result of conversational rule violations? *Journal of Psycholinguistic Research, 20,* 83–103.

Price, P. C., & Yates, J. F. (1995). Associative and rule-based accounts of cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1639–1655.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In H. Black, & W. K. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton–Century–Crofts.

Roberts, P. L., & MacLeod, C. (1995). Representational consequences of two modes of learning. *The Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 48A,* 296–319.

Shanks, D. R. (1990). Connectionism and the learning of probabilistic concepts. *The Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 42A,* 209–237.

Shanks, D. R. (1991a). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 433–443.

Shanks, D. R. (1991b). On similarities between causal judgements in experienced and described situations. *Psychological Science, 2,* 341–350.

Shanks, D. R. (1995). *The psychology of associative learning.* England: Cambridge University Press.

Shanks, D. R., & Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes, 24,* 405–415.

Shanks, D. R., & López, F. J. (1996). Causal order does not affect cue selection in human associative learning. *Memory & Cognition, 24,* 511–522.

Sherman, S. J., McMullen, M. N., & Gavanski, I. (1992). Natural sample spaces and the inversion of conditional judgments. *Journal of Experimental Social Psychology, 28,* 401–421.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119,* 3–22.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11,* 1–23.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88,* 135–170.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability learning. *Psychological Review, 90,* 293–315.

Wagner, A. R., Logan, F. A., Haberlandt, K., & Price, T. (1968). Stimulus selection in animal discrimination learning. *Journal of Experimental Psychology, 76,* 171–180.

Waldmann, M. R. (1996). Knowledge-base causal induction. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *Causal learning, the psychology of learning and motivation: Advances in research and theory* (Vol. 30, pp. 47–88). San Diego, CA: Academic Press.

Windschitl, P. D., & Weber, E. U. (1999). The interpretation of "likely" depends on the context, but "70%" is 70%-right? The influence of associative processes on perceived certainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1514–1533.

Wolford, G., & Taylor, H. A. (1990). The conjunction fallacy? *Memory & Cognition, 18,* 47–53.

Yates, J. F., & Carlson, B. W. (1986). Conjunction errors: Evidence for multiple judgement procedures, including "signed summation". *Organizational Behavior and Human Decision Processes, 37,* 230–253.

Appendix

Procedure Used to Simulate Participants' $p(D|S)$ Judgments in Experiment 1 With MDM

Memory traces consisted of vectors with six parts: one for each symptom and one for each disease. Each part had 12 components. The presence of a given symptom, or a given disease, was represented with the vector (1 1 1 1 −1 −1 −1 −1 0 0 0 0), whereas its absence was represented with the orthogonal vector (1 1 −1 −1 −1 −1 1 1 0 0 0 0). Thus, a patient with symptoms S_1 and S_2 suffering from R, for example, was represented with the following vector:

S_1 S_2
1 1 1 1 −1 −1 −1 −1 0 0 0 0 | 1 1 1 1 −1 −1 −1 −1 0 0 0 0 |

no S_3 no S_4
1 1 −1 −1 −1 −1 1 1 0 0 0 0 | 1 1 −1 −1 −1 −1 1 1 0 0 0 0 |

R no C
1 1 1 1 −1 −1 −1 −1 0 0 0 0 | 1 1 −1 −1 −1 −1 1 1 0 0 0 0 .

At the same time, each vector was conceived as formed by two minivectors: one for *S*, which included the first four parts for the symptom configuration, and another one for *D*, which included the last two parts for the diseases representation. This way, the information was represented to allow for $p(D|S)$ judgments. Memory traces were generated with the same program that generated the stimuli in each trial of Experiment 1. The program was run 160 times, each time including the 160 trials of the learning task. This yielded a total of 25,600 memory traces. With this number of traces, we could obtain stable simulation results.

To ask MDM to make $p(D|S)$ judgments, a symptom probe was constructed for each of the 15 symptom configurations shown in Table A1. For example, the symptom probe for the symptom configuration S_1 and no S_2 and no S_3 and no S_4 was as follows:

1 1 1 1 −1 −1 −1 −1 0 0 0 0|1 1 −1 −1 −1 −1 1 1 0 0 0 0|1 1
−1 −1 −1 −1 1 1 0 0 0 0|1 1 −1 −1 −1 −1 1 1 0 0 0 0|2 2 2 2
2 2 2 2 2 2 2 2|2 2 2 2 2 2 2 2 2 2 2 2.

The number 2 in a given component indicates that such a component does not enter into the similarity computation. Thus, only the *S* minivector of the symptom probe is relevant in this case. To ask a probability judgment conditional on a given symptom configuration, the *S* minivector of the symptom probe for that configuration was selected to probe memory in order to form the conditionalizing set. Then, the *S* minivector of all memory traces was compared with the *S* minivector of the probe, and only those whose similarity reached the similarity criterion (*Sc*) value were included into the conditionalizing set. Such similarity was calculated according to Dougherty et al.'s (1999) Equation 1.

Once the conditionalizing set was formed, a disease probe was constructed to compute the probability of each disease. For example, to compute the probability of Disease *R*, the following probe was constructed:

2 2 2 2 2 2 2 2 2 2 2 2|2 2 2 2 2 2 2 2 2 2 2 2|2 2 2 2 2 2 2 2 2 2
2 2|2 2 2 2 2 2 2 2 2 2 2 2|1 1 1 1 −1 −1 −1 −1 0 0 0 0|1 1 −1
−1 −1 −1 1 1 0 0 0 0.

In this case, only the *D* minivector of the probe was relevant. The *D* minivector of the probe was then compared with the *D* minivector of each trace in the conditionalizing set to calculate the conditional intensity according to Dougherty et al.'s (1999) Equation 6. After that, the probability of each disease was computed by normalizing the conditional intensity of each disease according to Dougherty et al.'s Equation 7.

We systematically varied *Sc* and the likelihood that each feature in every object was stored in memory (*L*) along the simulations to obtain the best fit measured as the sum of squared errors. The simulation data that best fit participants' judgments ($L = .8$ and $Sc = .7$) are shown in Table A1. The $SSE = 0.25$ was much larger than the *SSE* obtained with the neural network ($SSE = 0.04$). The simulation results do not improve if the absent symptoms in the probe are represented with a vector of 0s or with a vector of 2s.

We also used a normalization procedure based on the sigmoidal function. This was done by using the following equation:

$$p(D_R|S_j) = \frac{e^{I_c(D_R \cap S_j)}}{e^{I_c(D_R \cap S_j)} + e^{I_c(D_C \cap S_j)}},$$

where D_R and D_C are Diseases *R* and *C*, respectively, S_j is a symptom configuration, and I_c is the conditional intensity. Though this normalization procedure did improve MDM's fit ($L = 1$ and $Sc = 1$, $SSE = 0.10$), the sum of squared errors was still considerably larger than that obtained with the neural network. Finally, when no normalization procedure is used, the *SSE*s obtained from the simulations are also higher than 0.2.

Procedure Used to Simulate the Data From Experiment 3 With MDM

The procedure used to simulate the data from Experiment 3 with MDM was the same as for Experiment 1 except in the following respects. Memory traces were again generated by running the training program used for participants 160 times, each time including the 180 trials of the learning task. This yielded a total of 28,800 memory traces. The *S* minivector for each symptom probe was constructed by including the vector (1 1 −1 −1

Table A1
Summary of the Results Obtained From Participants' Judgments and the Simulated Data Obtained With the Neural Network and With MDM Experiment 1

	Symptom configurations														
Source	1000	0100	0010	0001	1100	1010	1001	0110	0101	0011	1110	1101	1011	0111	1111
P	.63	.45	.38	.15	.64	.64	.41	.39	.23	.14	.51	.34	.34	.21	.30
N	.63	.47	.38	.26	.61	.52	.39	.35	.24	.18	.49	.36	.28	.16	.25
MDM	.47	.28	.15	.06	.73	.51	.26	.23	.08	.03	.60	.35	.20	.06	.24

Note. In the vectors used to represent the symptom configurations, each component stands for the presence or absence of symptoms S_1 through S_4. The simulated data correspond to judgments about the probability of *R* given all symptom configurations except the no-symptoms one. P = participants' judgments; N = neural network; MDM = MINERVA-decision making.

(*Appendix continues*)

−1 −1 1 1 0 0 0 0) for each absent disease. We have not obtained better results using the alternative vector (2 2 2 2 2 2 2 2 2 2 2 2) or (0 0 0 0 0 0 0 0 0 0 0 0) for the absent symptoms. The *D* minivectors for the disease probe to compute the probability of *A*, *B*, and *A* and *B* were constructed, respectively, as follows:

A = 1 1 1 1 −1 −1 −1 −1 0 0 0 0 2 2 2 2 2 2 2 2 2 2 2 2,
B = 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 −1 −1 −1 −1 0 0 0 0, and
A and B = 1 1 1 1 −1 −1 −1 −1 0 0 0 0 1 1 1 1 −1 −1 −1 −1 0 0 0 0.

In this case, because the normalization procedure yielded very poor results, we did not use it to compute likelihood ratings. Thus, only Equation 6 (Dougherty et al., 1999) was used to compute probability ratings. A vector of 0s, instead of a vector of 2s, could also be used to represent the absence of information about a disease. However, this option yielded extremely poor results. Another alternative consists of taking the absence of information about a disease as the explicit absence of such disease. This means using the vector (1 1 −1 −1 −1 −1 1 1 0 0 0 0) for the absent disease. Not surprisingly, this also yielded a very poor quantitative fit (remember that participants' judgments were highly incompatible with such misunderstandings).

The simulation data that best fit participants' judgments (L = 1.0 and Sc = 0.7) are shown in Table 4. The quantitative fit of the model (SSE = 0.17) was substantially worse than that obtained with the double component (associative + averaging) model, SSE = 0.03.

Received February 7, 2001
Revision received June 20, 2002
Accepted July 1, 2002 ■