Journal of Behavioral Decision Making
J. Behav. Dec. Making, **12**: 275–305 (1999)

The 'Conjunction Fallacy' Revisited: How Intelligent Inferences Look Like Reasoning Errors

RALPH HERTWIG* and GERD GIGERENZER
Max Planck Institute for Human Development, Berlin, Germany

ABSTRACT

Findings in recent research on the 'conjunction fallacy' have been taken as evidence that our minds are not designed to work by the rules of probability. This conclusion springs from the idea that norms should be content-blind — in the present case, the assumption that sound reasoning requires following the conjunction rule of probability theory. But content-blind norms overlook some of the intelligent ways in which humans deal with uncertainty, for instance, when drawing semantic and pragmatic inferences. In a series of studies, we first show that people infer nonmathematical meanings of the polysemous term 'probability' in the classic Linda conjunction problem. We then demonstrate that one can design contexts in which people infer mathematical meanings of the term and are therefore more likely to conform to the conjunction rule. Finally, we report evidence that the term 'frequency' narrows the spectrum of possible interpretations of 'probability' down to its mathematical meanings, and that this fact — rather than the presence or absence of 'extensional cues' — accounts for the low proportion of violations of the conjunction rule when people are asked for frequency judgments. We conclude that a failure to recognize the human capacity for semantic and pragmatic inference can lead rational responses to be misclassified as fallacies. Copyright © 1999 John Wiley & Sons, Ltd.

KEY WORDS conjunction fallacy; probabalistic thinking; frequentistic thinking; probability

People's apparent failures to reason probabilistically in experimental contexts have raised serious concerns about our ability to reason rationally in real-world environments. One of the most celebrated of these failures is the *conjunction fallacy*, in which people violate what is widely considered the 'simplest and most fundamental qualitative law of probability' (Tversky and Kahnerman, 1983, p. 294), the *conjunction rule*. This rule holds that the mathematical probability of a conjoint hypothesis (A&B) cannot exceed that of either of its constituents, that is, $p(\text{A\&B}) \leqslant p(\text{A}), p(\text{B})$. Stich (1985), for instance, saw major implications of the conjunction fallacy for people's assessment of technological risks: 'It is disquieting to speculate on how large an impact this inferential failing may have on people's assessments of the chance of such catastrophes as nuclear reactor failures' (p. 119). Kanwisher (1989) argued that the conjunction fallacy might underlie 'flawed arguments' that 'often recur in debates on

* Correspondence to: Ralph Hertwig, Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: hertwig@mpib-berlin.mpg.de

CCC 0894–3257/99/040275–31$17.50

Accepted 23 November 1998

U.S. security policy ... Strategic priorities have in the past become distorted by overemphasizing the most extreme scenarios at the expense of less flashy but more likely ones' (pp. 652, 655). Finally, based on the conjunction fallacy, Gould (1992) concluded more generally that 'our minds are not built (for whatever reason) to work by the rules of probability' (p. 469).

Do violations of the conjunction rule justify such concern? In our view, this concern springs from the idea that norms should be content-blind — in the present case, the assumption that sound reasoning can be reduced to following the conjunction rule, and nothing else. We argue that content-blind norms overlook the intelligent ways in which humans deal with uncertainty, namely, the human capacity for semantic and pragmatic inferences. To illustrate this argument, let us consider the classical problem designed to demonstrate that people violate the conjunction rule: the Linda problem (Tversky and Kahneman, 1983). In this problem, participants read: 'Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.' Participants are then asked to rank several hypotheses according to their 'probability', including 'Linda is a bank teller' (T), 'Linda is active in the feminist movement' (F), and 'Linda is a bank teller and is active in the feminist movement' (T&F; or, in a variant, to indicate whether T or T&F is more 'probable').

Tversky and Kahneman (1983) concluded that the 80–90% of participants who judged T&F to be more probable than T were in error because they violated the conjunction rule. Applied to the Linda problem, the conjunction rule is a narrow norm in two senses (Gigerenzer, 1996). First, the norm is applied in a content-blind way, that is, it assumes that judgments about what counts as sound reasoning may ignore content and context (Gigerenzer and Murray, 1987, Chapter 5). For instance, in Kahneman and Tversky's (1996) view of sound reasoning, the content of the Linda problem is irrelevant; one does not even need to read the description of Linda. All that counts are the terms 'probability' and 'and', which the conjunction rule interprets as mathematical probability and logical AND, respectively (Gigerenzer and Regier, 1996; Hertwig, 1995). In addition, probability theory is imposed as a norm for a single event (whether Linda is a bank teller). This would be considered misguided by those statisticians who hold that probability theory is about repeated events (for details see Gigerenzer, 1994).

The account we propose is a step toward integrating content, context, and representation of information. It puts the burden of explanation on cognitive processes that infer the meaning of natural language terms, rather than on laws of reasoning that may or may not conform to the conjunction rule. In doing so, it contributes to the arguments of several authors that content and pragmatic context are indeed relevant from the perspective of conversational principles in the Linda problem and other conjunction problems (e.g. Adler, 1984, 1991; Dulany and Hilton, 1991; Hilton, 1995; Politzer and Noveck, 1991).

In addition, our account contributes to an explanation of the empirical finding that representing problems in terms of absolute frequencies rather than probabilities largely reduces the conjunction fallacy, an effect first described by Tversky and Kahneman (1983). Whereas Kahneman and Tversky (1996; Tversky and Kahneman, 1983) attributed this *frequency effect* to 'extensional cues' in frequency representations that facilitate reasoning according to the conjunction rule (henceforth, *extensional-cue hypothesis*), we propose that a major cause of the frequency effect is that a frequency representation is less polysemous than a probability representation (henceforth, *semantic-inference hypothesis*).

SEMANTIC INFERENCE AND POLYSEMY

Many natural language terms are *polysemous*, that is, they have multiple meanings that are linked. In this article, we use the term *semantic inference* for the process by which a mind infers the particular

meaning of a polysemous term from the semantic content and the pragmatic context of the sentence, paragraph, or text in which it occurs. In the Linda problem, participants are not informed that the problem is an exercise in probability or set theory. Therefore, they have to make an inference about the meaning of the instruction, in particular about what the experimenter means by asking them to judge 'probability' (for related arguments see Macdonald, 1986; Margolis, 1987; Teigen, 1994, p. 227).

Is the term 'probability' polysemous? In the view of Ian Hacking (1975), the notion of mathematical probability was two-faced from its very beginning in the seventeenth century. Its aleatory face was concerned with observed frequencies (e.g. co-occurrences of comets and kings' deaths); its epistemic face was concerned with degrees of belief or opinion warranted by authority. Barbara Shapiro (1983) and Lorraine Daston (1988) have argued that the term 'probability' actually had more than two faces in the seventeenth and eighteenth centuries. Meanings also included physical propensity (e.g. as inferred from the physically symmetrical construction of a die), frequency, strength of argument (e.g. evidence for or against a judicial verdict), intensity of belief (e.g. the firmness of a judge's conviction of the guilt of the accused), verisimilitude, and epistemological modesty. The advent of the mathematical theory of probability eliminated many of these meanings but left at least three: degree of belief, objective frequency, and physical propensity (Gigerenzer *et al*., 1989). Unlike the mathematical theory of probability, however, the natural language concept of probability has retained many of its earliest meanings. A glance at the entries for 'probability' and 'probable' in a few dictionaries illustrates this.

The *Oxford English Dictionary* (1989) lists the following contemporary meanings of 'probability': 'the appearance of truth, or likelihood of being realized, which any statement or event bears in the light of present evidence', 'something which, judged by present evidence, is likely to be true, to exist, or to happen', and 'likelihood'. The *Third New International Dictionary* (1967) links 'probable' to the existence of 'evidence'. It also lists 'credible', as in 'a credible story', among the meanings of 'likely'. The *Random House College Thesaurus* (1984) associates 'probable' with 'presumable, expected, possible, credible, logical, reasonable, plausible, tenable, thinkable, conceivable, believable'.

To resolve the polysemy of 'probability' and 'probable', one must draw semantic inferences from the context in which the terms are used.[1] It is evident that most of the candidate meanings of 'probability' and 'probable' cannot be reduced to mathematical probability. For instance, if one interprets 'probability' and 'probable' in the Linda problem as 'something which, judged by present evidence, is likely to happen', 'plausible', or 'a credible story', then one might easily judge T&F to be more probable than T because the information about Linda was deliberately selected to provide no evidence for the hypothesis 'bank teller'. Under these interpretations it is pointless to compare participants' judgments with a norm from mathematical probability theory, because the inferred meanings have nothing to do with mathematical probability.

In German, the language spoken by the participants in the experiments reported below, the term equivalent to probable is *wahrscheinlich*, which, like 'probable', is polysemous. The *Duden* (1981), a German dictionary and the primary authority on matters of the German language, reports that the term *wahrscheinlich* derives from the Latin word *verisimilis*, a combination of *verus* (true) and *similis* (similar). 'Similar' and 'having a resemblance' were also former meanings of the English term 'likely', according to the *OED*. This reveals an interesting relation between the etymology of 'probable' in German and English and the representativeness heuristic, which some researchers have argued underlies participants' probability judgments in conjunction problems. In the Linda problem, application of

[1] There is a rich research tradition on how people map linguistic probability terms onto numerical equivalents (for an excellent review, see Budescu and Wallsten, 1995). We, in contrast, focus on how the single term 'probability' is mapped onto its mathematical and nonmathematical meanings.

the representativeness heuristic has been characterized as a 'similarity' or 'typicality' assessment (Shafir, Smith and Osherson, 1990; Smith and Osherson, 1989; Tversky and Kahneman, 1983).

SEMANTIC INFERENCE BY SOCIAL RATIONALITY

The mind has to decide which of the various meanings of the term 'probability' to apply in any given context. Which of the various meanings of the terms 'probability' and 'probable' do participants infer in the Linda problem? One answer can be derived from Paul Grice's (1975, 1989) theory of conversational reasoning, a form of social rationality. His theory's fundamental assumption is that it is reasonable for the audience to assume that the communicator is trying to meet certain general standards of communication. From knowledge of these standards, observation of the communicator's behavior, and the context, the audience should be able to infer the communicator's specific informative intention.

In the Linda problem, one of these standards — the *relevance* maxim (originally termed the *maxim of relation* by Grice), according to which the audience expects the communicator's contribution to be relevant — allows one to make the following prediction. Participants who assume that the relevance maxim applies to the Linda problem should infer that 'probability' does not mean mathematical probability (e.g. frequencies), because a mathematical interpretation would render the experimenter's description of Linda (i.e. the communicator's contribution) irrelevant to the requested judgment (Adler, 1984, 1991).

To preserve the relevance maxim, then, participants must infer that alternative meanings of 'probability' and 'probable', such as 'something which, judged by present evidence, is likely to happen', or 'plausible', apply, so that the participant is required to assess the description of Linda. We refer to these meanings as *nonmathematical* as opposed to *mathematical* (e.g. frequency, percentage). If one of these nonmathematical meanings is inferred, then choosing T&F is not a violation of probability theory because mathematical probability is not being assessed. We propose that semantic inferences about the meaning of 'probability' that fall outside of mathematical probability strongly contribute to why many people violate the conjunction rule in the Linda problem.

PREDICTIONS

The thesis is that human minds resolve the uncertainty in the Linda problem by intelligent semantic inferences that exploit social rationality, here conversational maxims in human interaction. This reinterpretation of the conjunction effect leads to three predictions about inferred meanings of 'probability', 'frequency', and 'believability', and the proportion of *conjunction violations* that will be observed when participants are asked to make each of these judgments in the Linda problem (or similar problems); we use the term conjunction violation to denote a judgment inconsistent with the conjunction rule.

- *Prediction 1*: Probability judgments. If asked for probability judgments, people will infer a nonmathematical meaning of 'probability', and the proportion of conjunction violations will be high as a result. This prediction directly follows from the assumption that the relevance maxim renders mathematical meanings of 'probability' implausible and favors other relevance-preserving interpretations that cannot be reduced to mathematical probability. Prediction 1 has an interesting implication. It should be possible to construct a situation in which people following conversational maxims are less likely to violate the conjunction rule. For instance, if immediately before the

probability judgment participants are asked for a judgment that renders the description of Linda relevant — such as a typicality judgment (i.e. 'How good an example of a bank teller is Linda?') — then the proportion of conjunction violations in a subsequent probability judgment will decrease. The reason is that semantic inferences about the meaning of 'probability' are no longer constrained by the maxim of relevance (as the information about Linda has already been used in the same conversational context). Furthermore, the *maxim of quantity* makes a nonmathematical interpretation less plausible (for the detailed argument, see below). Prediction 1 and its implication will be tested in Studies 1 and 2.

- *Prediction 2*: Frequency judgments. If asked for frequency judgments, people will infer mathematical meanings, and the proportion of conjunction violations will decrease as a result. This prediction derives from the assumption that the term 'frequency', unlike 'probability', narrows down the spectrum of possible interpretations to meanings that follow mathematical probability. Thus, mathematical meanings will be inferred *despite* the assumption that the relevance maxim favors relevance-preserving interpretations. Prediction 2 also has an interesting implication. The presence or absence of 'extensional cues' in the frequency representation (such as specifying the numerical size of a reference class, Tversky and Kahneman, 1983) should have little effect provided that the term 'frequency' is used. Prediction 2 and its implication will be tested in Studies 3 and 4.
- *Prediction 3*: Believability judgments. If the term 'probability' is replaced by 'believability', then the proportion of conjunction violations should be about as prevalent as in the probability judgment. The relevance maxim renders mathematical probability an implausible interpretation of 'probability' and favors nonmathematical interpretations. Because 'believability' is one of the nonmathematical and relevance-preserving interpretations of 'probability', replacing one by the other should yield a similar proportion of judgments that do not conform to the conjunction rule. This prediction differs from that of Macdonald and Gilhooly (1990), who proposed that using the term 'probability' (as opposed to 'believability') might 'cue formal probabilistic rules' (p. 60) and thus decrease conjunction violations.

STUDY 1: WHAT MEANINGS OF 'PROBABILITY' DO PEOPLE INFER IN THE LINDA PROBLEM?

Prediction 1 states that if asked for probability judgments, people will infer a nonmathematical meaning of 'probability', and the proportion of conjunction violations will be high as a result. How can one access people's semantic inferences in the Linda problem? We used two approaches. First, we asked participants to paraphrase the term 'probability' in the context of the Linda problem (*paraphrase* task). Second, we gave them a list of meanings of 'probability' and asked them to check off which corresponded best to their understanding of the term (*choice* task).

Method

Participants were first presented with the Linda problem and instructed to rank the two constituent hypotheses and the conjoint hypothesis (T, F, T&F) according to their probability. Order of hypotheses was counterbalanced across participants. The precise wording of the probability representation is displayed later in Exhibit 3. They were then asked to imagine themselves in the role of an experimenter who has to describe the Linda problem verbally to a fictitious participant who is not a native speaker of German. The instructions stated that the term 'probability' exceeds the fictitious participant's linguistic competence and must be paraphrased. Participants' oral paraphrases of the Linda problem were audio recorded. Then participants listened to their own taped descriptions and

were asked to turn their paraphrases of 'probability' into precise instructions for the fictitious participant; this was done to ensure that we understood the paraphrases correctly.

Following the paraphrase task, the same participants were given a list of 13 different interpretations of the term 'probability' and were asked to check off the one(s) that best reflected their understanding of it in the Linda problem. The list consisted of the terms 'logicality', 'certainty', 'frequency', 'typicality', 'credibility', 'plausibility', 'tenability', 'evidence', 'conceivability', 'possibility', 'predictability', 'similarity', and 'reasonableness'. Most of the interpretations were included because they were generated in a pilot study of the paraphrase task. Although the representativeness heuristic does not address the issue of semantic inferences, it assumes that people judge probability in the Linda problem by assessing similarity (Tversky and Kahneman, 1983). Therefore, if one were to derive a prediction from the representativeness account, one would expect to find evidence that participants infer that probability means 'similarity' or 'typicality'. To keep track of such inferences, we included them in the list as well.

In addition to these 13 interpretations, participants were given the option to add their own interpretation to the list, if it was not already included. Each interpretation was presented in terms of hypothetical problem instructions, for instance, 'judge the possibility that Linda is a bank teller, is active in the feminist movement, is a bank teller and is active in the feminist movement'. The interpretations were presented in two different random orders.

Participants

Eighteen students at the University of Munich recruited by advertisement from a broad spectrum of disciplines served as participants. None had previously encountered the Linda problem, and each was tested individually.

Results

A conjunction violation is defined here as a judgment in which the ranked or estimated probability (frequency) of a conjoint hypothesis (e.g. T & F) exceeds the probability (frequency) of the constituent hypothesis (e.g. T, F). A paraphrase is counted as mathematical if it reflects one of the main mathematical interpretations of probability, which include, according to the *Cambridge Dictionary of Philosophy* (1995), classical ('expectancy'), relative frequency ('frequency', 'percentage'), logical ('logicality'), and subjective ('certainty') interpretations. Other paraphrases are classified as nonmathematical. Some participants produced adjectives rather than nouns as paraphrases of 'probability'; for the purpose of comparison, we transformed adjective paraphrases into appropriate nouns.

Conjunction violations

Consistent with previous results (e.g. Teigen, Martinussen and Lund, 1996), 15 of the 18 participants (83%) violated the conjunction rule in the Linda problem. Did this happen because nonmathematical meanings of 'probability' are inferred under the relevance maxim (Prediction 1)? To the extent that participants' paraphrases reflect their understanding of 'probability', we can answer this question.

Paraphrases

Eighteen participants produced 39 responses (on average, 2.2 each). Bearing testimony to the polysemy of 'probability', the 39 responses include 18 different interpretations. Exhibit 1(a) shows the frequency of different paraphrases. Only 7 of the 39 (18%) responses were mathematical (including 'randomness',

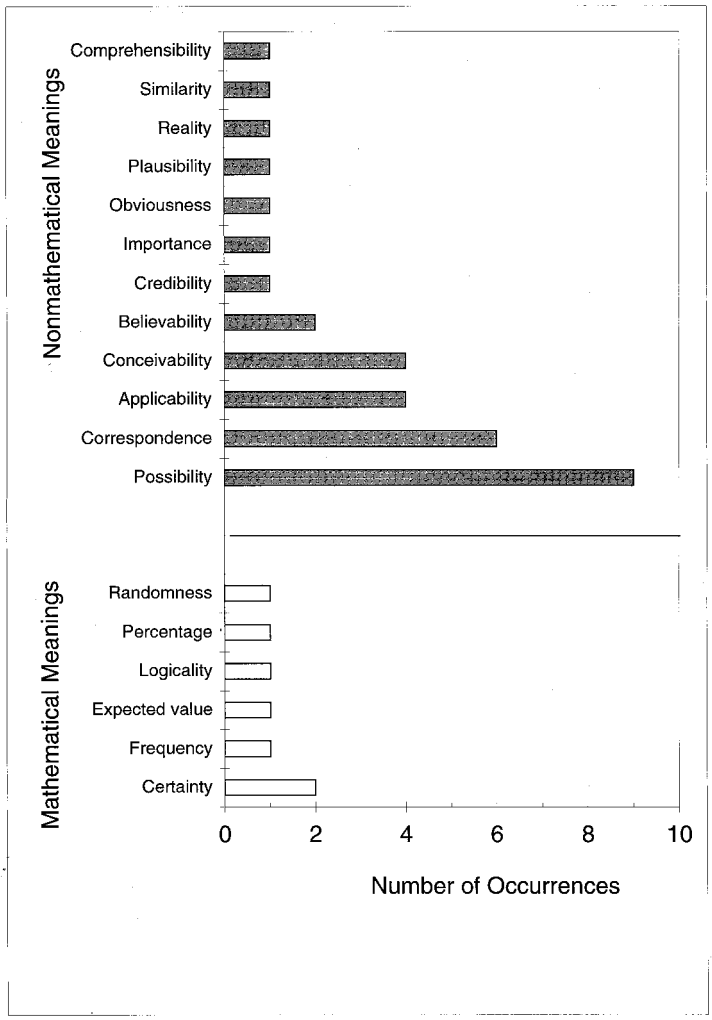

Exhibit 1(a). Frequency of mathematical and nonmathematical paraphrases of 'probability'

which is not explicitly mentioned in the *Cambridge Dictionary of Philosophy*), while the great majority were nonmathematical. Of the three participants who did not violate the conjunction rule, two produced at least one mathematical meaning, whereas 12 of the 15 participants who violated the conjunction rule did not produce a single mathematical paraphrase.

The four most frequent paraphrases were 'possibility' ($n = 9$), 'correspondence' ($n = 6$), 'applicability' ($n = 4$), and 'conceivability' ($n = 4$). Unlike the other three paraphrases, 'correspondence' is a heterogeneous category that includes instructions such as 'Judge the correspondence between Linda's world view and her activities', and 'Judge the correspondence between Linda's characteristics and her activities'. One might argue that 'correspondence' can be related to 'similarity' and thus provides evidence for the use of the representativeness heuristic. Even if one combines 'correspondence' and 'similarity' into one category, only 7 of the 32 (22%) nonmathematical responses could be interpreted as evidence for the representativeness account.

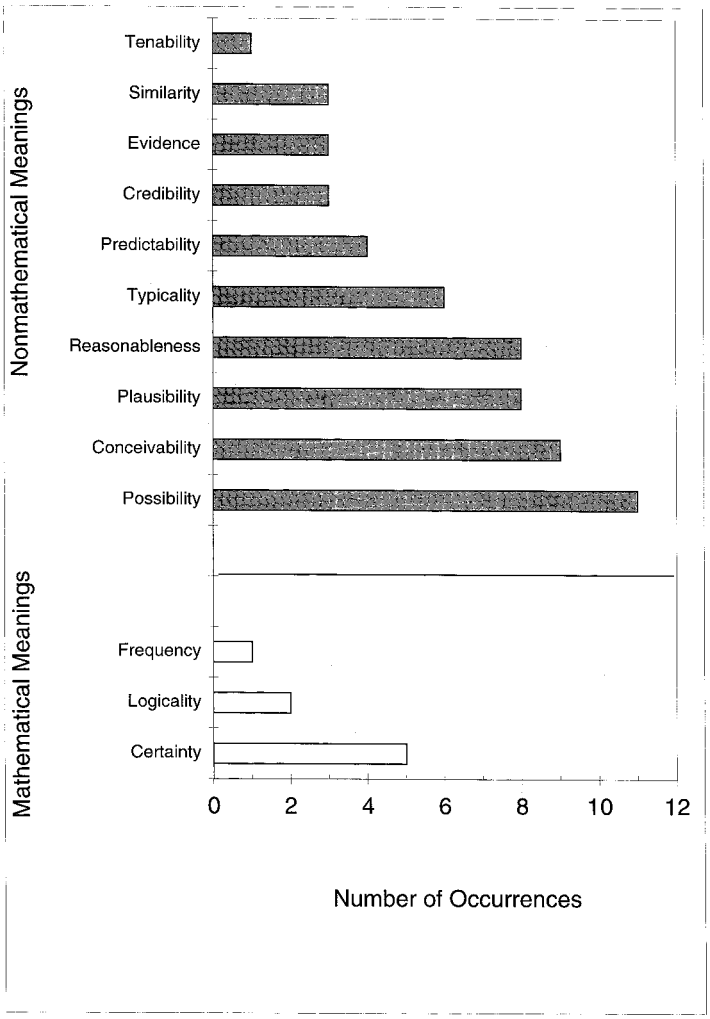

Exhibit 1(b). Frequency with which mathematical and nonmathematical interpretations of 'probability' were chosen

Choices

The 18 participants made a total of 64 choices (on average, 3.6 each). Exhibit 1(b) shows how often each interpretation was chosen. Consistent with Prediction 1, most choices were nonmathematical (88%, 56 of 64). The four most frequent interpretations were 'possibility' ($n = 11$), 'conceivability' ($n = 9$), 'plausibility' ($n = 8$), and 'reasonableness' ($n = 8$). Of the 56 nonmathematical choices, 'typicality' and 'similarity' were chosen a total of 9 times (16%). Note that the list of 13 different interpretations included more nonmathematical than mathematical meanings because it was based on interpretations that were generated in a pilot study of the paraphrase task. Could the imbalance explain why the majority of participants chose nonmathematical meanings? To control for this possibility, we calculated the average number of choices separately for mathematical and nonmathematical meanings. On average, each of the mathematical interpretations was chosen 2.7 times, whereas each of the

nonmathematical meanings was chosen more than twice as often (5.6 times). That is, the fact that the list included more nonmathematical than mathematical meanings cannot explain why nonmathematical meanings were chosen more frequently.

Summary
Study 1 tested whether people infer nonmathematical meanings of 'probability' in the Linda problem. The results are consistent with Prediction 1. Whether participants were asked to paraphrase the term 'probability' or to choose the meaning that best reflected their understanding of the term, the majority of them strongly favored nonmathematical meanings. This result raises the interesting possibility of designing a context in which the relevance maxim can be preserved without inferring nonmathematical meanings. In the next study, we propose one way to construct such a context.

STUDY 2: HOW TO PRESERVE THE RELEVANCE MAXIM

In Study 2, participants were first asked for a typicality judgment and then for a probability judgment in the Linda problem. This procedure allowed participants to use the description of Linda before the probability judgment, thus preserving the relevance maxim (which says that the audience expects the communicator's contribution to be relevant). As a consequence, participants were left free to infer other than relevance-preserving meanings in the subsequent probability judgment.

What meanings of 'probability' will participants infer when this judgment is preceded by a typicality judgment? Another of Grice's (1975) conversational maxims, namely, the maxim of quantity ('Make your contribution as informative as is required') leads to the prediction that the interpretation of 'probability' as mathematical probability (e.g. relative frequency, percentage) will now be a possible inference, and that other interpretations — such as 'possibility' or 'conceivability' — will be less likely to be inferred. Here is the rationale for this. The description of Linda was intentionally designed (Tversky and Kahneman, 1983, p. 297) so that the response to the typicality question will likely be: 'She is a very good example of a feminist, a less good example of a feminist bank teller, and a rather poor example of a bank teller.' If in the same conversational context the question 'Which is most probable?' is then posed, interpretations of 'probable' (e.g. 'possible') that lead to a judgment identical to the typicality judgment (i.e. $p(\mathrm{F}) > p(\mathrm{T\&F}) > p(\mathrm{T})$) become less plausible, because such a ranking has just been given, and thus to repeat it would be uninformative. In other words, because asking the same question twice and providing the same answer twice in the same conversational context would not be cooperative, participants are now entitled to assume that the probability judgment is different from the typicality judgment (Hilton, 1995).[2] If the maxim of quantity favors mathematical meanings, then the proportion of conjunction violations will decrease.

To obtain evidence for this prediction in addition to the judgments, we asked participants to think aloud while they were working on both the typicality and probability judgments. If nonmathematical meanings become less plausible when participants make a typicality judgment first, then the think-aloud protocols should reveal cognitive processes indicative of mathematical interpretations.

[2] The conversational context may also help in another respect. Bless, Strack and Schwarz (1993) pointed out that adjacent questions in social discourse should be related to each other and thus suggested that 'as adjacent questions normally refer to each other, it is very likely that subjects try to use the content of the preceding question to disambiguate the meaning of an ambiguous subsequent one' (p. 154, see also Hilton, 1995).

Method

Participants were randomly assigned to one of four conditions. In the TP^+ condition, participants first gave the typicality (T) judgment and then the probability (P) judgment. They were instructed to think aloud (+) as they made both judgments. Participants received the same probability representation of the Linda problem as in Study 1 (Exhibit 3). The P^+ condition, in which participants were asked to think aloud but made only the probability judgment, served as a control for the TP^+ condition. In the TP condition, participants made the typicality judgment and the probability judgment and were not asked to think aloud. As a control for the TP condition, we added the P condition, in which participants made only the probability judgment. Participants familiarized themselves with the think-aloud procedure by working on two practice problems. Their protocols were audio recorded and later transcribed. Participants in the P^+, TP, and TP^+ conditions were tested individually.

In the TP^+ and TP conditions, participants first received a booklet with, on the first page, the description of Linda, followed by a joint introduction to the typicality and probability judgments that read: 'Now we would like to ask you two questions concerning Linda.' The next page showed an instruction adapted from Rosch (1975):

> The first question has to do with what we have in mind when we use words which refer to specific categories. For instance, take the word dog. You certainly have some notion of what a 'real dog' is. For many people, a St Bernard is a dog which is more typical of the category dog than a Pekinese. Notice that this has nothing to do with how well you like St Bernards or Pekinese. Possibly you like Pekinese better than St Bernards but recognize the St Bernard to be more representative of the category dog. The judgment of *how good an example of a category various instances of the category are* can be made in different areas. You will be asked to make such a judgment. We will ask you to judge the degree to which Linda, who is described on the preceding page, corresponds to the ideas you have of certain categories.

On the next page, the hypotheses *Linda is an active feminist*, *Linda is a bank teller*, and *Linda is a bank teller and an active feminist* were listed, each accompanied by a 7-point rating scale. Participants were instructed to rate the correspondence between Linda and their idea or image of these categories. Below the typicality task on the same page, the hypotheses were stated again and participants were asked: 'Which of the following statements is most probable? Assign that statement a ranking of 1, the second most probable statement a ranking of 2, and the remaining statement a ranking of 3.'

Participants

Ninety students at the University of Salzburg recruited by advertisement from a broad spectrum of disciplines served as participants ($n = 22$ in each of the TP^+, P^+, and TP conditions and 24 in the P condition). None of the participants had previously encountered the Linda problem.

Results

One implication of Prediction 1 is that conjunction violations will decrease if one asks participants for a typicality judgment immediately before the probability judgment. In the P and P^+ conditions respectively, 88% (21 of 24) and 82% (18 of 22) of the participants violated the conjunction rule, whereas in the TP^+ condition, 27% (6 of 22) did so, and in the TP condition 59% (13 of 22). Thus, the addition of a typicality judgment immediately before the probability judgment decreased the proportion of conjunction violations, whether or not the judgments were accompanied by a think-aloud instruction (P versus TP: a difference of 29 percentage points; P^+ versus TP^+: a difference of 55 percentage points). According to Cohen (1988), these effects are of medium to large effect size

($\phi = 0.32$, $\phi = 0.55$). The effect of the typicality judgment was larger when participants thought aloud than when they did not. We do not have an explanation for this difference. Comparison of the P and P^+ conditions indicates that thinking aloud *per se* (and thus also time spent thinking about the task) hardly affects the proportion of conjunction violations (88% versus 82%).

Analysis of the think-aloud protocols

Think-aloud protocols might reveal the cognitive processes associated with the specific semantic inferences participants made. We expected to see more evidence of processes related to mathematical interpretations of probability in the TP^+ than in the P^+ condition. Indeed, 75% (12 of 16) of the participants in the TP^+ condition who ranked T&F least probable either mentioned the principle of class inclusion (5 of 12) or considered the combination bank teller and feminist to result in an empty set due to their incompatibility (7 of 12). In contrast, 67% (12 of 18) of the participants in the P^+ condition who ranked T least probable argued that some aspect of the description of Linda (e.g. majored in philosophy) is incompatible with being a bank teller. Of the four participants who ranked T&F least probable in the P^+ condition, one referred to class inclusion, while two others considered T&F to be an empty set.

Summary

Most previous studies that found a high proportion of violations of the conjunction rule used a probability representation such as that used in Study 2. Applying Grice's (1975, 1989) theory of conversational reasoning to the Linda problem, we successfully created a context in which the proportion of conjunction violations decreased in the probability representation of the Linda problem. The think-aloud data indicated that most participants in the TP^+ condition argued in terms of logic (sets) and probability theory (class inclusion) to justify ranking T&F least probable. In contrast, participants' reasoning in the P^+ condition seemed to be related to nonmathematical interpretations of 'probability' (lack of correspondence between Linda and the category 'bank teller').

STUDY 3: WHAT MEANINGS OF 'FREQUENCY' DO PEOPLE INFER IN THE LINDA PROBLEM?

Prediction 2 states that, if asked for frequency judgments, people will infer mathematical meanings and the proportion of conjunction violations will decrease as a result. We know of two studies that used *frequency representations* (as opposed to *probability representations*) of conjunction problems, and two studies that presented frequency information but asked for a probability and possibly a percentage judgment. The exact wording of the problems used in three of these four studies can be found in the Appendix. In the first of these studies, Tversky and Kahneman (1983) contrasted probability and frequency representations in the health-survey problem. Fiedler (1988) used probability and frequency representations of seven problems, including the Linda problem. Reeves and Lockhart (1993) compared what they called 'frequency' problems and 'case-specific' problems. Finally, Jones, Jones and Frisch (1995) examined two problems, including the Linda problem, in what they called 'frequency' and 'single-case' versions, without providing their precise wording.

The results for the frequency problems in the above studies varied. In the health-survey problem, Tversky and Kahneman found 58% and 25% conjunction violations in the probability and frequency representations, respectively. Fiedler (1988, Experiment 1) reported 73% conjunction violations in the probability representation and 23% in the frequency representation, averaged across all seven

problems. A second experiment with five problems yielded similar results (means of 70% and 25% in the probability and frequency representations, respectively). Averaged across the problems in Experiment 1, Reeves and Lockhart (1993) found 59% conjunction violations for probability problems and 35% for frequency problems.[3] Finally, Jones *et al.* (1995) reported an average of 85% conjunction violations (across two problems) in probability problems and 81% in frequency problems. The decrease in the proportion of conjunction violations from probability to frequency problems in these studies ranged all the way from 4% in Jones *et al.* (1995) to 50% in Fiedler (1988).

Why are these results so varied? We argue that the 'frequency' problems are not the same across studies. They differ in a number of ways, making it impossible to determine the factor(s) that make people reason in accordance with the conjunction rule more often in one study than in another (see Appendix). For instance, Tversky and Kahneman (1983) did not describe a particular individual in the health-survey problem, only a reference class of 100 adult males, whereas Fiedler described a particular individual (e.g. Linda) and introduced a reference class of 100 people. Reeves and Lockhart (1993) did not even ask for frequency judgments but rather provided frequency evidence on which probability judgments were to be based, and it is not clear whether Jones *et al.* (1995) asked for frequency or percentage judgments (see pp. 112, 113).

Explanations of the frequency effect

How have these previous researchers explained the frequency effect? As already mentioned, Tversky and Kahneman (1983; Kahneman and Tversky, 1996) attributed it to the presence of extensional cues in the frequency representation. The extensional cue they suggested was responsible for the frequency effect was 'an explicit reference to the number of individual cases', which 'encourages subjects to set up a representation of the problems in which class inclusion is readily perceived and appreciated' (Tversky and Kahneman, 1983, p. 309). Fiedler (1988) attributed the effect to polysemy: 'The prevailing statistical interpretation of probability (as relative frequency) does not appear to apply to colloquial language' (p. 123). Reeves and Lockhart (1993) attributed the effect to two different ways in which people derive probabilities if the evidence at hand is frequentistic versus based on a single case. These different approaches, which they call 'singular' versus 'distributional', in turn 'evoke different kinds of inferential rules and heuristic procedures, some of which are more closely aligned with extensional principles than others' (p. 207). Finally, Jones *et al.* (1995) did not provide an explanation for the frequency effect.

One factor that none of the previous authors suggested but that may contribute to the frequency effect is the difference in response mode. The frequency representation asks for quantitative estimates, whereas the probability representation typically asks for ranks. Could this difference have contributed to the frequency effect? A growing body of evidence suggests that it could. In a series of three studies, involving 460 participants (students at German and American universities and lay people), Hertwig and Chase (1998) found that participants faced with the Linda problem are more likely to follow the conjunction rule in probability estimates than in rankings — the overall difference in the number of conjunction violations totaled 30 percentage points (medium effect: $\phi = 0.3$). Hertwig and Chase attributed this finding to the fact that people apply combination rules in estimation but not in ranking (for their detailed account, see pp. 324–329 in Hertwig and Chase, 1998).

Does this finding extend to the frequency representation, that is, do more people follow the conjunction rule in frequency estimates than in rankings? We know of only one study that asked participants to both rank and estimate frequencies in the Linda and other conjunction problems.

[3] Because Reeves and Lockhart (1993) varied problem representation within participants, the results reported are for the representation that participants encountered first.

Consistent with Hertwig and Chase's finding, Erdfelder, Bröder and Brandt's (1998) results showed that people are more likely to follow the conjunction rule in frequency estimates than in rankings, the difference in the Linda problem being about 30 percentage points (Erdfelder *et al.*'s Study 2).

This 30-point difference, however, still accounts for less than half of the frequency effect. The frequency effect describes the fact that the frequency representation yields fewer conjunction violations than the probability representation. For instance, in the Linda problem, this difference is about 70 percentage points.[4] About half of this difference appears to be accounted for by the response mode. For instance, if the estimation response is replaced by a ranking response, conjunction violations increase by about 30 percentage points in the frequency representation. Even then, however, a frequency ranking representation yields about 30–40 percentage points fewer conjunction violations than a probability ranking representation. The semantic-inference hypothesis suggests that this effect is due to the replacement of 'probability' by 'frequency'.

In Study 3, we test whether the term 'frequency', unlike 'probability', in fact narrows down the possible spectrum of meanings to those that follow mathematical probability. In Study 4, we attempt to isolate the factor or factors, other than response mode, that are responsible for the frequency effect. In particular, we investigate whether the effect is due to the replacement of 'probability' by 'frequency', as the semantic-inference hypothesis holds. This hypothesis is contrasted with the extensional-cue hypothesis, which holds that extensional cues facilitate reasoning according to the conjunction rule.

Method

Using a procedure similar to that in Study 1, in Study 3 we asked participants to paraphrase the term 'frequency'. Participants first were presented with the Linda problem in a frequency representation and instructed to estimate the frequency of the two constituent hypotheses and the conjoint hypothesis (T, F, T&F). The order of the hypotheses was counterbalanced across participants. The precise wording of the frequency representation is displayed in Exhibit 3. Following the frequency judgment, participants received the same instruction as in the paraphrase task of Study 1, except that they were asked to paraphrase the term 'frequency'.

Participants

Twenty students at the University of Munich recruited by advertisement from a broad spectrum of disciplines served as participants. None of the participants had previously encountered the Linda problem, and each participant was tested individually.

Results

Mathematical meanings of 'frequency' are defined as paraphrases that are directly related to numerosity (e.g. 'proportion', 'number', etc.); other paraphrases that do not refer to numerosity (e.g. 'plausibility') are classified as nonmathematical meanings. Twenty participants produced 55 responses (on average, 2.8 each), encompassing 11 different interpretations. (However, some interpretations were closely related, e.g. 'proportion' and 'percentage', 'expectancy value' and 'favorable/possible' events,

[4] This difference can be determined as follows. Hertwig and Chase (1998) reviewed a sample of 17 conditions in 10 studies in which the proportion of conjunction violations in the probability ranking representation of the Linda problem was examined and found a median of 87%. The median proportion of conjunction violations in the frequency estimation representation of the Linda problem in the studies reviewed here is 17% (22%, 17% in Experiments 1 and 2 of Fiedler, 1988; 85% in Experiment 1 of Jones *et al.*, 1995, and 0% and 13% in our Studies 3 and 4). Thus, the median proportion of violations for the probability ranking representation is 70 percentage points higher than that for the frequency estimation representation.

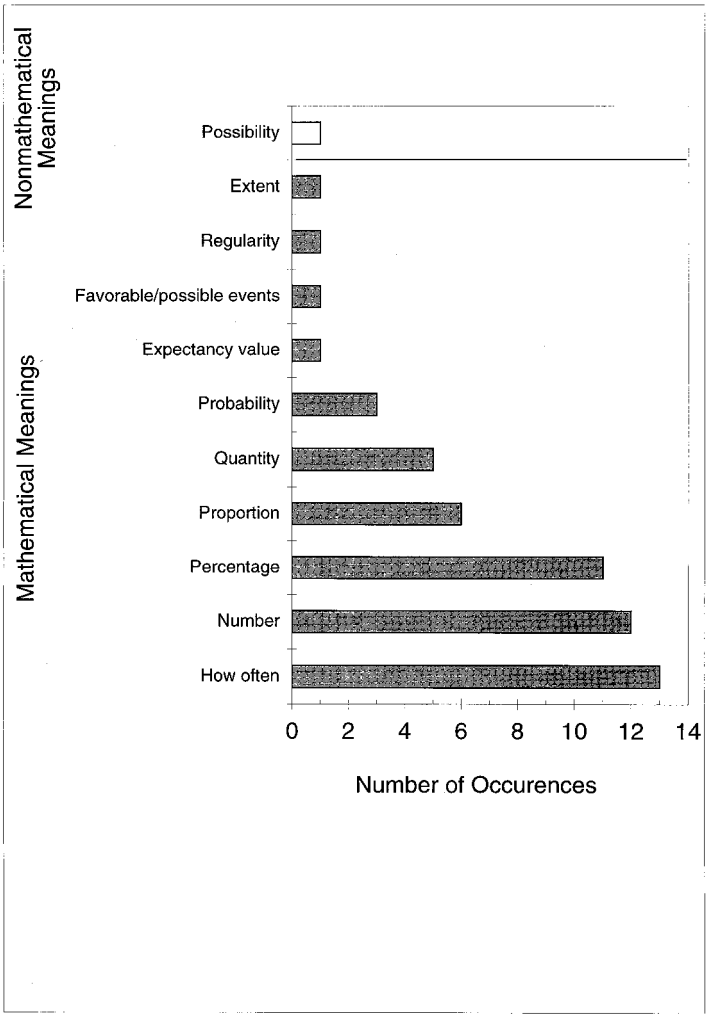

Exhibit 2. Frequency of mathematical and nonmathematical paraphrases of 'frequency'

or 'number' and 'how often') Exhibit 2 shows the frequency of different paraphrases; only one of the 55 responses ('possibility') was nonmathematical. The five most frequent paraphrases were 'how often'[5] ($n = 13$), 'number' ($n = 12$), 'percentage' ($n = 11$), 'proportion' ($n = 6$), and 'quantity' ($n = 5$).

Studies 1, 2, and 3 allow for a comparison of the prevalence of conjunction violations for probability and frequency judgments. When participants were asked for probability judgments in Studies 1 and 2 (with or without the think-aloud instruction), fewer than one in five followed the conjunction rule. When asked for frequency judgments in Study 3, participants never violated the conjunction rule. In Study 4, we test the implication of Prediction 2 that the presence or absence of 'extensional cues' in the frequency representation should have little effect.

[5] 'How often' is a smart solution to the paraphrase instruction that we did not foresee. It is smart insofar as this term mimics the phrase 'how many', which is already provided in the frequency representation ('How many of the 200 women are bank tellers?').

STUDY 4: WHAT IS THE IMPACT OF EXTENSIONAL CUES?

Tversky and Kahneman (1983) mentioned only one extensional cue, the specification of the size of the reference class, that might be responsible for the frequency effect. The health-survey problem provides another feature that might count as an extensional cue. In the probability representation of the health-survey problem, the scanty information given (i.e. adult males in British Columbia of all ages and occupations, see Appendix) refers to a class of people and links features to a reference class. In contrast, in the probability representation of the Linda problem, features (e.g. 'single', 'outspoken') are linked to an individual. If the linking of features to a reference class as opposed to an individual is a cue to the inclusion relation between sets, then one could in fact explain why Tversky and Kahneman found a smaller conjunction effect in the health-survey than in the Linda problem (58% versus 80–90%). To examine this possibility, we tested as possible extensional cues the linking of features to a reference class and specification of the size of the reference class.

The procedure was as follows. To determine a standard against which variants could be compared, we first tested a frequency representation of the Linda problem that includes all extensional cues; this frequency representation is identical to that tested in Study 3. The frequency representation and all of its variants are shown in Exhibit 3. Then, in Test 1, we constructed variants of the frequency representation in which we eliminated the first and then both of the two candidate extensional cues, to find out whether the extensional-cue hypothesis or the semantic-inference hypothesis accounts for the frequency effect. In the first variant, the features are linked to a single person rather than to a reference class of 200 women. In the second variant, the size of the reference class is left unspecified (see Exhibit 3). Both variants used the term 'frequency'. If the extensional-cue hypothesis is correct, then the deletion of one of these cues (according to Tversky and Kahneman, 1983, the specification of the size of the reference class) or both cues should eliminate the frequency effect. If the semantic-inference hypothesis, which holds that replacing 'probability' by 'frequency' is crucial, is correct, then the deletion of one or both of these cues should not eliminate the frequency effect. All variants used the term 'frequency'.

In Test 2, we designed a probability variant of the Linda problem that mimics the frequency representation as closely as possible without asking for a frequency judgment (see Exhibit 3). This variant includes the two 'extensional cues', that is, it specifies the size of a reference class and links the features to it. In addition, it asks for estimates instead of ranks. This variant mimics the frequency representation in every way except for the feature that the semantic-inference hypothesis considers critical: It includes the polysemous term 'probability'. Thus, the extensional-cue hypothesis predicts that the proportion of conjunction violations should strongly decrease in this variant. In contrast, the semantic-inference hypothesis predicts that the proportion of conjunction violations should not decrease relative to that in the estimation response mode.

Method

Exhibit 3 lists the frequency representation of the Linda problem and three variants. Participants were randomly assigned to one of the four resulting Linda problems. All problems presented the two constituent hypotheses (e.g. T, F) and the conjoint hypothesis (e.g. T&F). Order of hypotheses was counterbalanced across participants.[6]

[6] In this experiment, some of the participants also received the health-survey problem and the Boris Becker problem (the latter an update of Tversky and Kahneman's, 1983, Björn Borg problem). These two problems were always presented after the Linda problem. We will return to these two problems in the discussion when we address the question of why the conjunction effect has been much stronger in the Linda problem than in problems such as health-survey and Boris Becker.

Participants
Participants were 98 students at the University of Salzburg recruited through advertisement from a broad spectrum of disciplines. Participants were tested in groups of six or less. None had previously encountered the Linda problem.

Results

Frequency representation
Replicating the results in Study 3, we found that only 13% of the participants (3 of 23; one participant did not complete the problem) violated the conjunction rule in the frequency representation of the Linda problem. This result and those that follow are shown in Exhibit 3.

Test 1. What in the frequency representation matters?
Does the linking of the features matter? In the probability representation of the Linda problem, a single individual is described; in the frequency representation, the features are linked to a reference class. In one variant, we changed this factor in the frequency representation by relating the features to a single individual while leaving everything else the same. The reference class was introduced after the description (see Exhibit 3). Note that this variant is close to the frequency representation used by Fiedler (1988), who also introduced a reference class after Linda's description. The proportion of conjunction violations, 20% (5 of the 25 participants who received the Linda problem only), matched Fiedler's (1988) 22%. The proportion of conjunction violations we found for this variant was only 7 percentage points higher than for the frequency representation, and 68 percentage points ($\phi = 0.68$) lower than for the probability representation tested in Study 2 (P condition: 88%, 21 of 24). Consistent with the semantic-inference hypothesis, the linking of the features (either to a single individual or to a reference class) cannot account for the frequency effect.

Does the specification of the size of the reference class matter? In the second variant, in addition to describing a single individual, the size of the reference class was left unspecified (see Exhibit 3). This variant, in which participants had to generate the size of the reference class on their own, puts Tversky and Kahneman's (1983) suggestion that an explicit reference to the number of individual cases accounts for the frequency effect to an empirical test. In this variant, only 16% (4 of 25) of the participants violated the conjunction rule. This number is only 3 percentage points higher than for the frequency representation, and 72 percentage points ($\phi = 0.72$) lower than for the probability representation tested in Study 2. Consistent with the semantic-inference hypothesis, the reference to the number of individual cases cannot account for the frequency effect. The implication of Prediction 2 thus passed Test 1.

Test 2. Does adding extensional cues matter?
In Test 1, the strategy was to eliminate first one and the both extensional cues to determine whether either of these manipulations could make the frequency effect disappear. In Test 2, we started from the other direction by using the probability representation and adding both extensional cues. This probability variant links features to the reference class and provides the size of the reference class (see Exhibit 3). It differs from the frequency representation only in asking for a probability rather than a frequency judgment. If the semantic-inference hypothesis is correct, then this variant should elicit the large proportion of conjunction violations observed in the probability representation because it still employs the polysemous term 'probability'.

Exhibit 3. Frequency and probability representations of the Linda problem and variants thereof (note that here only one constituent hypothesis, T, is listed; participants read hypotheses T, F, and T&F)

Representation	Study/proportion of conjunction violations
Probability representation Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice and also participated in anti-nuclear demonstrations. Rank the following hypotheses according to their probability. Linda is a bank teller.	*Study 1*: 83% (15 out of 18) *Study 2: P condition*: 88% (21 out of 24) *Study 2: P^+ condition*: 82% (18 out of 22)
Frequency representation[a] In an opinion poll, the 200 women selected to participate have the following features in common: They are, on average, 30 years old, single, and very bright. They majored in philosophy. As students, they were deeply concerned with issues of discrimination and social justice and also participated in anti-nuclear demonstrations. Please estimate the frequency of the following events. How many of the 200 women are bank tellers? ____ of 200	*Study 3*: 0% (0 out of 20) *Study 4*: 13% (3 out of 23)
Frequency variant: single person features Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. [...] Imagine 200 women who fit the description of Linda. Please estimate the frequency of the following events. How many of the 200 women are bank tellers? ____ of 200	*Study 4, Test 1*: 20% (5 out of 25)
Frequency variant: single person features and without specification of the reference class size Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. [...] Imagine women who fit the description of Linda. Please estimate the frequency of the following events. How many of these women are bank tellers? ____ of ____	*Study 4, Test 1*: 16% (4 out of 25)
Probability variant with extensional cues In an opinion poll, the 200 women selected to participate have the following features in common: [...] From these 200 women, a single person, Linda, was selected by chance. Please estimate the probability of the following events. Linda is a bank teller.	*Study 4, Test 2*: 67% (16 out of 24)

[a]Fiedler (1988) reported that the frequency effect is the same whether or not a round number (e.g. 100 versus 168) is chosen for the reference class. Thus, for simplicity, we chose a round number.

The proportion of conjunction violations for this probability variant, which has all extensional cues, was 67% (16 of 24). This result is 54 percentage points ($\phi = 0.55$) higher than for the frequency representation, and 21 percentage points ($\phi = 0.25$) lower than for the probability representation tested in Study 2. We attribute this difference to the fact that participants in Test 2 were asked to estimate rather than to rank probabilities (Hertwig and Chase, 1998). We can also compare this variant to the two frequency variants examined in Test 1, both of which used the term 'frequency' and an estimation response mode but lacked one and two of the extensional cues, respectively. Relative to the result for the frequency variant lacking one extensional cue, the proportion of conjunction violations for the present variant is 47 percentage points higher ($\phi = 0.47$). Relative to the result for the frequency variant lacking two extensional cues, it is 51 percentage points higher ($\phi = 0.52$).

Thus, Test 2 provided further evidence that, aside from response mode, the single most important factor for reasoning in accord with the conjunction rule is the substitution of 'probability' by 'frequency'. The implication of Prediction 2 thus passed Test 2.

Summary

Prediction 2 stated that, if asked for frequency judgments, people will infer mathematical meanings and the proportion of conjunction violations will decrease. Study 3 provided evidence that people in fact infer mathematical meanings in the frequency representation. The frequency representations tested in Studies 3 and 4 provided independent evidence that the proportion of conjunction violations decreased dramatically in a frequency representation.

Prediction 2 has a strong and testable implication: The presence or absence of 'extensional cues' in the frequency representation (such a specifying the numerical size of a reference class, Tversky and Kahneman, 1983) should have little effect, provided that the term 'frequency' is used. The results for two variants of the frequency representation and one variant of the probability representation support the implication: The presence of the two extensional cues did not make a majority of people reason according to the conjunction rule, and their absence did not appreciably increase the proportion of conjunction violations.

BELIEVABILITY JUDGMENTS

Prediction 3 states that if the term 'probability' is replaced by 'believability', then the proportion of conjunction violations should be about the same as in the probability judgment. We have tested this prediction using data reported by Macdonald and Gilhooly (1990), who tested the hypothesis that probability judgments increase inclusion judgments relative to believability judgments because they cue formal probabilistic rules. The semantic-inference hypothesis makes a different prediction: Because 'believability' is one of the relevance-preserving interpretations of 'probability', replacing the one term by the other should leave the judgments largely unaffected.

In Macdonald and Gilhooly's (1990) design, one group of participants read a Linda problem in which the instruction was 'Rank the following statements according to their probability, using 1 for the most probable and 8 for the least probable'. A second group received the same instruction, except that 'probability' and 'probable' were replaced by 'believability' and 'believable'. Macdonald and Gilhooly (1990) reported the proportion of conjunction violations for the probability judgments (89%), but not for the believability judgments; they did, however, report the mean ranks of T, F, and T&F. For this kind of data, Prediction 3 forecasts similar mean ranks for both kinds of judgments. For the probability judgments, the mean ranks were 5.9, 2.0, and 4.6 for T, F, and T&F, respectively, and for the believability judgments, the mean ranks were 5.7, 1.9, and 4.8 for T, F, and T&F, respectively. Consistent with Prediction 3, the ranks for probability and believability judgments are similar.

Summary and discussion

Recent research on the conjunction fallacy has demonstrated that human probability judgments may violate the conjunction rule. In problems such as the Linda problem, we argue that semantic inferences about the meaning of the polysemous term 'probability' contribute to this finding. Study 1 showed that in the Linda problem most paraphrases of 'probability' are nonmathematical. Participants' inferences about the meaning of the term seem to be guided by the maxim of relevance, a tool of social rationality. Study 2 showed that when semantic inferences about the meaning of 'probability' are no longer constrained by the maxim of relevance and nonmathematical meanings are less plausible because of the maxim of quantity, the proportion of conjunction violations decreases. Using a frequency representation is another way to direct such semantic inferences. Study 3 showed that people's paraphrases of the term 'frequency' in the Linda problem refer almost exclusively to numerosity.

With Study 4 we aimed to determine which specific properties of a frequency representation make people reason according to the conjunction rule. The results for three variants of the frequency and probability representations of the Linda problem tested showed that the use of the term 'frequency' rather than 'probability' is one of the two most important factors behind the frequency effect, the other being response mode (Erdfelder *et al*., 1998; Hertwig and Chase, 1998). The results of Study 4 help to explain why mixed results have been obtained in the few studies that have looked at reasoning according to the conjunction rule using frequency representations. The small effect obtained by Reeves and Lockhart (1993), for instance, can easily be explained by the fact that what they called 'frequency' problems (see Appendix) instructed participants to assess 'probabilities'. Prediction 3 states that replacing 'probability' by 'believability' will leave the proportion of conjunction violations unchanged. Results reported by Macdonald and Gilhooly (1990) support this prediction.

OBJECTIONS

We can think of several objections to the semantic-inference hypothesis. All of them hinge on the argument that one can find the conjunction effect even under conditions where the semantic-inference hypothesis appears to predict adherence to the conjunction rule. We address each in turn.

The willingness-to-bet instruction

In their original article, Tversky and Kahneman (1983) considered the possibility that violations of class inclusion 'may be viewed as a misunderstanding regarding the meaning of the word *probability*' (p. 303). To test this possibility, they examined a problem in which the term 'probability' was dropped. Although we do not agree that interpretations of 'probable' other than mathematical probability should be labeled 'misunderstandings', this test does address the issue of polysemy. The problem read:

> Consider a regular six-sided die with four green faces and two red faces. The die will be rolled 20 times and the sequence of greens (G) and reds (R) will be recorded. You are asked to select one sequence, from a set of three, and you will win $25 if the sequence you chose appears on successive rolls of the die. Please check the sequence of greens and reds on which you prefer to bet.
>
> 1. RGRRR
> 2. GRGRRR
> 3. GRRRRR

In two groups, 65% and 62% of participants, respectively, ranked sequence 2 highest, although sequence 2 includes sequence 1. From this result Tversky and Kahneman (1983) concluded that

'our subjects followed the representativeness heuristic even when the word [probability] was not mentioned' (p. 303). This result is surprising given that Tversky and Kahneman (1983, p. 300) also tested a betting version of the Linda problem and found a result consistent with the semantic-inference hypothesis we propose. In this test, participants were given Linda's description and the instruction: 'If you could win $10 by betting on an event, which of the following would you choose to bet on?' With this betting instruction, conjunction violations decreased by 33 percentage points relative to the standard problem (from 89% to 56%).

Why does the betting instruction decrease the proportion of conjunction violations in the Linda problem but not in the die problem? Although the die problem includes the polysemous term 'probability', unlike the Linda problem, it provides little information that would be rendered irrelevant by a mathematical interpretation of 'probability'. This may explain why the replacement of 'probability' does not affect the proportion of conjunction violations. However, the die problem has other ambiguities that participants need to resolve, and which may account for the responses of the majority of participants who violated the conjunction rule with the betting instruction. For instance, Macdonald (1986) argued that participants may infer that the task is to bet on the numbers rather than the sequences of reds and greens, in which case 'sequence (2) is by far the most likely' (p. 22; by the binomial formula, the probabilities of sequences 1, 2, and 3 are 0.041, 0.082, and 0.016, respectively). In this context, it is also interesting to note that the betting representation of the die problem contained an extensional cue (i.e. an explicit reference to the number of cases, namely, 20 rolls), which Kahneman and Tversky (1996) suggested is responsible for the low proportions of conjunction violations in frequency problems.

Bar-Hillel and Neter (1993) also used a 'willingness-to-bet' instruction in three studies in which they explored violations of the disjunction rule, according to which the probability of A-or-B cannot be smaller than the probability of either A or B. In two of the three studies, Bar-Hillel and Neter found that the willingness-to-bet instruction led to a small average increase in judgments consistent with the disjunction rule compared with the probability instruction. (In Study 1, for instance, judgments consistent with the disjunction rule increased in 8 of 10 problems by 2 to 18 percentage points.) As in the die problem, we suggest that in these problems there are other ambiguities that participants have to resolve, although Bar-Hillel and Neter were aware of the importance of conversational maxims and took some trouble not to violate them. To illustrate one ambiguity that we see as crucial, consider the following problem, which is representative of those used by Bar-Hillel and Neter:

> [Dorit] writes letter home describing an English-speaking country, where a modern Western technological society coexists with primal landscape and free-roaming wildlife. Where was the letter written?

Participants received the following four options (among others), two of them representative (South Africa, Africa) and two of them unrepresentative (Holland, Europe). Participants had to rank the places according to their probability. The content-blind disjunction rule requires, for instance, that Africa be ranked over South Africa. However, such a response violates the maximum of quantity, as it is not as informative as required by the instructions, which assert a level of specificity, namely, an 'English-speaking country' that is not met by the answer 'Africa' (i.e. a continent, not a country).

We suggest that the conversational goal of being informative may have contributed to the finding that a majority of participants violated the disjunction rule with both the 'probability' and the 'willingness-to-bet' instruction. This might also explain why violations also occurred when they should not have according to a representativeness explanation: Bar-Hillel and Neter (1993, p. 1124) proposed that if a set and a subset are unrepresentative, then representativeness predicts a preference for the set over the subset. Nevertheless, in the Dorit problem, for instance, 43% and 37% of participants ranked

the unrepresentative subset 'Holland' over the unrepresentative set 'Europe' in the probability task and betting task, respectively. Again, if participants were trying to match the level of specificity implied by the instructions in their responses, this result is not surprising.

Clarification of the problem's statistical nature

To compensate for the 'misleading communication; (p. 6) in the Linda problem, Donovan and Epstein (1997) instructed participants as follows: 'Please consider all of the problems that follow as essentially statistical problems. Some of the problems test your ability to find a disguised statistical problem. Your job is to avoid being distracted by extraneous information or by the form in which the problem is presented' (p. 9). This instruction addresses the semantic-inference hypothesis in that it may make nonmathematical meanings of 'probability' less likely to be inferred.

Did the instruction increase the proportion of inclusion judgments? Based on their results, Donovan and Epstein (1997) concluded: 'The essence of the difficulty of the notorious Linda problem cannot be attributed to the violation of implicit conversational rules' (p. 16). Despite this conclusion, and consistent with the semantic-inference hypothesis, however, across all conditions Donovan and Epstein observed proportions of violations about 30 percentage points lower than the median of inclusion judgments typically found in the Linda problem (see footnote 4).

Frequency judgments

The seven-letter problem is a conjunction problem for which Tversky and Kahneman (1983, p. 295) reported violations of class inclusion in frequency estimates. In this problem, most participants estimated the number of seven-letter words of the form '____ing' to be higher than the number of seven-letter words of the form '_____n_'. It is unclear whether this result was found in a within-subjects or in a between-subjects design (see the conflicting statements in Tversky and Koehler, 1994, p. 547 and Kahneman and Tversky, 1996, p. 586). If it was in a within-subject design, then the effect may have resulted from not presenting the alternatives simultaneously. (We return to the issue of between-subjects designs in the next section.) We therefore hypothesized that when both alternatives are presented to the same group of participants simultaneously, this effect will largely disappear.

To test this hypothesis, we studied the following seven-letter problem both in a within-subject and in a between-subjects design:

> In four pages of a novel (about 2000 words), how many words would you expect to find that have the form ____ing (that is, seven-letter words that end with 'ing') and how many words have the form _____n_ (that is, seven-letter words that have the letter n in the sixth position)? Estimate the frequency of these words.
> _____ of the 2000 have the form ____ing
> _____ of the 2000 have the form _____n_.

One group of participants ($n = 31$) received both options. Two other groups of participants judged only one of the two alternatives ($n = 31$ per group). Consistent with our conjecture, participants in the between-subjects design estimated on average that 425 (median = 300) of the 2000 words end with 'ing', and that 212 (median = 50) of the 2000 words have the letter n in the sixth position. In contrast, participants in the within-subject design estimated on average that 282 (median = 150) of the 2000 words end with 'ing', and that 457 (median = 250) of the 2000 words have the letter n in the sixth position. In this condition, 26% (8 of 31) of the participants violated class inclusion.

In sum, we found that few violations of internal consistency occur if participants see both alternatives. If participants are provided with only one alternative, however, their judgments violate the inclusion rule. The difference between within-subjects and between-subjects designs is important for more than the seven-letter problem. For instance, Birnbaum and Mellers (1983) found that research on the 'base-rate fallacy' leads to different conclusions depending on whether a within-subjects or a between-subjects design is used: 'The interpretation of base rate ''neglect'' is based on the finding that in between-subjects comparisons, the effect of base rate is too small ... In within-subject comparisons, however, subjects use the base rate, and the evidence for a ''fallacy'' disappears' (Varey, Mellers, and Birnbaum, 1990, p. 623). Birnbaum (1982) gave a range-frequency analysis of why between-subjects comparisons lead to paradoxical conclusions: They confound the stimulus and the context by allowing the stimulus to evoke its own context. For these reasons, we agree with Varey *et al.* (1990), who argued that 'one should be extremely cautious when drawing inferences from between-subjects comparisons of judgments' (p. 623), particularly when one aims to investigate adherence to a rule of *internal* consistency.

Between-subjects design

Tversky and Kahneman (1983) used both within- and between-subjects designs to examine the Linda problem. In the latter design, participants either judged the critical constituents (i.e. T, F) or their conjunction (T&F), not both, as in the within-subjects design. In both designs, the probability of p(T&F) was ranked as higher than p(T). In the between-subjects design, however, a mathematical interpretation of 'probability' does not appear to render the experimenter's description of Linda irrelevant to the requested judgment because no participant judges both T and T&F. Is this finding therefore inconsistent with the semantic-inference hypothesis? We do not think so, for the following reason.

As already illustrated, the natural language term 'probability' is polysemous. The polysemy holds regardless of whether or not a participant gets to see all the critical hypotheses in the Linda problem. In both designs, participants have to decide which of the various meanings of the term 'probability' to infer. So far, we have intentionally limited our treatment of semantic inference to conversational maxims (i.e. the relevance maxim and the quantity maxim), as they allowed us to construct a situation in which people following the maxims are less likely to violate the conjunction rule (see also the following discussion on the health-survey problem). Semantic inference, however, is likely to be guided by other cues as well. In fact, the study by Donovan and Epstein (1997) mentioned earlier indicates that these inferences can be guided by the knowledge that the Linda problem is meant to be a statistical problem.

Several other candidate cues can be added to the list. Semantic inferences concerning the term 'probability' in reasoning problems may also be directed by (1) whether or not some sort of statistical information, such as base rates or likelihoods, is provided (e.g. in Bayesian reasoning problems); (2) whether or not the problem provides a clear sampling space and sampling process and describes the influence of chance in producing events (e.g. in sample-size problems; see Nisbett *et al.*, 1983); (3) whether or not solving the problem requires computations (as, for instance, in Bayesian reasoning problems and conditional probability problems); and finally, (4) the reference class from which the problem is drawn (e.g. ball-and-urn problems may steer semantic inferences more directly toward mathematical interpretations of 'probability' than do problems involving assessment of personality descriptions).

In the between-subjects design of the Linda problem, all of these (possible) cues point toward nonmathematical meanings of 'probability'. In solving the problem, participants are usually *not* told that this is an exercise in probability theory, are *not* provided with any statistical information (only

individuating information), are *not* required to make computations of any sort, and so on. As a consequence, a mathematical interpretation of 'probability' is not likely to be the default in the context of the Linda problem — even if such an interpretation would not render the experimenter's description of Linda irrelevant to the requested judgment, as in the between-subjects design. Likewise, one would not expect a mathematical interpretation of 'probability' to be the default interpretation if class inclusion is less salient in the Linda problem (as, for instance, in Tversky and Kahneman's, 1983, 'subtle' tests) and the violation of the relevance maxim therefore less obvious. For similar reasons, we are unconvinced by the argument that because the information about Linda is relevant for the ordering of T and F (but not T and T&F, or F and T&F), mathematical meanings of 'probability' should be inferred in the Linda problem. As Sperber and Wilson (1995, p. 123) pointed out, 'relevance is a matter of degree'; mathematical probability is not the interpretation of 'probability' that maximizes the relevance of the information provided in the Linda problem.

One point is very clear: We need a better understanding of the processes underlying semantic inferences, and we are far from having a process model that could capture those inferences (although important steps in this direction have been made; e.g. Sperber and Wilson, 1995). Modeling these processes is particularly important in a research field that relies as heavily on text problems to study cognitive processes as we do in research on probabilistic reasoning.

WHAT ARE OTHER WAYS OF PRESERVING THE RELEVANCE MAXIM?

From the perspective of a conversational analysis, the crux of the Linda problem is that the 'initial personality sketch would be an uncooperative, because irrelevant, contribution under a purely formal reading' (Adler, 1991, p. 255). Having not been informed that the relevance maxim is suspended, the participant tries to preserve the relevance maxim. In Study 1, we identified one way in which the participant can do this — by inferring meanings of 'probability' that require an assessment of Linda's description. However, this is not the only way to preserve the relevance maxim. Based on a conversational analysis, several researchers (e.g. Adler, 1984, 1991; Dulany and Hilton, 1991; Politzer and Noveck, 1991) argued that participants interpret the hypothesis 'Linda is a bank teller' to mean 'Linda is a bank teller and is not active in the feminist movement' (T&not-F implicature). On such a reading, Linda's personality sketch is relevant to deciding between T&F and T&not-F.

Different routes have been taken to block this implicature. Tversky and Kahneman (1983, p. 299) replaced 'Linda is a bank teller' by 'Linda is a bank teller *whether or not she is active in the feminist movement*'. This formulation led to a marked decrease in the proportion of conjunction violations (from 89% to 57%). Messer and Griggs (1993) used a similar clarifying phrase ('Linda is a bank teller, regardless of whether or not she is also active in the feminist movement') and observed a decrease in the proportion of conjunction violations from 77% to 56%. Morier and Borgida (1984) inserted the hypothesis 'Linda is a bank teller who is not a feminist' into the original set and found 77% conjunction violations, which barely differed from results obtained in the baseline probability estimation version (80%). However, in another conjunction problem, the Bill problem (see Tversky and Kahneman, 1983, p. 297), they found a decrease in the proportion of conjunction violations from 77% (baseline) to 49% (clarified hypothesis set).

Agnoli and Krantz (1989) compared two rephrasings of the conjunction problem — one designed to make the T&not-F implicature explicit and the other to block it — to a standard phrasing. In the first rephrasing, the hypothesis 'Linda is a bank teller' was replaced by 'Linda is a bank teller and not a feminist' (not-F group); in the second rephrasing, it was replaced by 'Linda is bank teller and may or may not be a feminist' (may-be-F group). According to Agnoli and Krantz (1989), if participants draw the T&not-F implicature, then the incidence of violations will be the same for the standard phrasing

and the not-F group, and fewer violations will occur in the may-be-F group (see critical discussion of Agnoli and Krantz's rationale in Politzer and Noveck, 1991). None of the predictions that Agnoli and Krantz (1989) attributed to this hypothesis have been corroborated.

Macdonald and Gilhooly (1990) tried to block the T&not-F implicature by rephrasing 'who is active in the feminist movement' as 'who may or may not be active in the feminist movement' and required participants to predict what 'will most probably be true of Linda in ten years'. In contrast to Agnoli and Krantz (1989), Macdonald and Gilhooly (1990) found that the proportion of conjunction violations decreased from 75% in the standard phrasing to 21% in the rephrasing. Politzer and Noveck (1991) also attempted to block the T&not-F implicature by employing an event E (*Roland took an exam*) that logically entails two possible outcomes, O_1 (*Roland failed an exam*) and O_2 (*Roland passed an exam*). They predicted that in the judgment of hypotheses E, O_1, and O_2, no implicature would be drawn and thus the incidence of conjunction violations would decrease. Indeed, they obtained 29% conjunction violations in this problem, compared to 77% in the control condition (in which participants solved the Linda problem).

Dulany and Hilton (1991, Experiment 1) asked participants who had already completed the Linda problem to identify which interpretation of the bank-teller hypothesis corresponded most closely to the one they inferred. Four possible interpretations were offered: Linda is a bank teller — (1) and is not active, (2) and is probably active, (3) and is probably not active, and (4) whether or not she is active — in the feminist movement. They argued that only those participants who chose interpretation (4) and nevertheless did not follow the conjunction rule can be said to have actually violated the rule. Of 61 participants, 32 (52%) violated the conjunction rule; 16 of them chose interpretations 1–3, while 16 others selected the formal reading. Thus, by Dulany and Hilton's criterion, only 26% of the participants can be said to have violated the conjunction rule.

With few exceptions, the evidence reported so far indicates that at least some participants preserve the relevance maxim by drawing the T&not-F implicature: Estimates range from 20% to 50% of participants. In our studies, we focused on another way to preserve the relevance maxim, namely, by inferring meanings of 'probability' that require the participant to assess the description of Linda. These two different ways of preserving the relevance maxim are not exclusive (and probably also not exhaustive). For instance, when we asked participants in Study 1 to respond to questions concerning their understanding of the bank teller hypothesis, 35% of them (6 of 17; 1 participant did not respond) indicated they had interpreted T to mean T&not-F. We also found that all participants who drew the Tnot-F implicature inferred at least one nonmathematical meaning of 'probability' in the paraphrase task of Study 1.

We argue that using the conjunction rule as a content-blind norm for sound reasoning means overlooking the capacity of the human mind to draw these relevance-preserving inferences. Rather than assuming that there is a one-to-one correspondence between natural language and probability theory, we propose modeling these inferences, which are drawn under uncertainty.

HEALTH-SURVEY VERSUS LINDA PROBLEM

There is a puzzling phenomenon in the literature on the conjunction fallacy. Problems in different content domains elicit different proportions of conjunction violations. Although Tversky and Kahneman (1983) classified problems by content according to whether they included 'representative' or 'causal' conjunctions (for research on causal conjunctions, see e.g. Ahn and Bailenson, 1996; Fabre, Caverni and Jungermann, 1995; Thüring and Jungermann, 1990), the representativeness explanation does not predict different proportions of conjunction violations for the two types of contents. The semantic-inference hypothesis offers an explanation of why the Linda problem, which includes a

representative conjunction, leads to higher proportions of conjunction violations than the health-survey and the Boris Becker problems, which include causal conjunctions.

We argue that it is in the Linda problem where social rationality, here embodied by conversational maxims, conflicts most with a logical reading of the problem. Unlike the Linda problem, the health-survey and Boris Becker problems provide hardly any individuating information about the people in question, Mr F and Boris Becker. Thus, although all three problems involve the polysemous term 'probability', in the health-survey and the Boris Becker problems there is little information that a mathematical interpretation of 'probability' could render irrelevant. Moreover, the reference classes from which these problems are drawn, medicine and sport, may steer semantic inferences about 'probability' more directly toward mathematical interpretations than the Linda problem, which involves assessment of personality descriptions. These speculations based on the semantic-inference hypothesis have two testable implications. First, because mathematical interpretations of 'probability' are more plausible in the health-survey and Boris Becker problems, the proportions of conjunction violations in these problems should be lower than in the Linda problem. Second, because a frequency representation should discourage nonmathematical interpretations of probability equally across problems, the semantic-inference hypothesis predicts low and comparable proportions of conjunction violations in all three problems in this representation.

We tested frequency and probability representations of the health-survey and Boris Becker problems as well as the Linda problem. Consistent with the implication above, 63% (15 out of 24) and 58% (14 out of 24) of the participants who received the probability representation of the health-survey and Boris Becker problems violated the conjunction rule, whereas 88% of the participants in the Linda problem did so (Study 2, P condition; for a similar result, see also Tversky and Kahneman, 1983, pp. 302, 305). That is, the proportion of conjunction violations we found in the Linda problem was 25 percentage points ($\phi = 0.29$) and 30 percentage points ($\phi = 0.33$) higher than in the health-survey problem and Boris Becker problem, respectively. Of the participants who received the frequency representation, in contrast, 13% (3 out of 24) in the health-survey problem, 9% (2 out of 23) in the Boris Becker problem, and 13% (3 out of 23, see Study 4) in the Linda problem violated the conjunction rule.

HOW DO FREQUENCY REPRESENTATIONS IMPROVE STATISTICAL REASONING?

Most supposed cognitive illusions in probabilistic reasoning, such as base-rate neglect and the conjunction fallacy, have been demonstrated using problems represented in terms of probabilities. The conjunction fallacy is not the only cognitive illusions that is reduced, or even made to disappear, when participants are given frequency information and asked for frequency judgments instead of probability judgments. For example, Gigerenzer (1994; Gigerenzer, Hoffrage and Kleinbölting, 1991) showed that the 'overconfidence bias' disappears when participants estimate the number of correct answers instead of the probability that a particular answer is correct (see also May, 1987; Sniezek and Buckley, 1993; Juslin, Olsson and Winman, 1998). Koehler, Gibbs and Hogarth (1994) reported that the 'illusion of control' (Langer, 1975) is reduced when the single-event format is replaced by a frequency format, that is, when participants judge a series of events rather than a single event. Finally, Bayesian reasoning improves in lay people (Cosmides and Tooby, 1996; Gigerenzer and Hoffrage, 1995) and experts (Hoffrage and Gigerenzer, 1998) when Bayesian problems are presented in natural frequencies (i.e. absolute frequencies obtained by natural sampling) rather than in a single-event probability format. Natural frequencies have also proven very effective in training people how to make conjoint and conditional probability judgments and Bayesian inferences (Sedlmeier, 1997).

These results should not be interpreted to mean that frequency judgments are always accurate. To the extent that we succeed in modeling the mechanisms underlying frequency and probability judgments, we will be able to predict when people's frequency judgments are valid and invalid, according to certain norms, and to explain why. There already exist precise explanations for when and why frequency judgments make overconfidence disappear and improve Bayesian reasoning. According to the theory of probabilistic mental models, judgments of confidence in each of a series of single events (i.e. that a given answer is correct) and a judgment of the frequency of correct answers across events can differ because they refer to different kinds of reference classes (Gigerenzer *et al*., 1991). Gigerenzer and Hoffrage (1995) showed than when numerical information is represented by natural frequencies rather than by relative frequencies or probabilities, then Bayesian reasoning involves fewer steps of mental computation.

To these explanations we can add another. In this article we have shown that the frequency effect in research on the conjunction effect can be at least partly explained by people's social rationality. People bring to the experimental context a repertoire of assumptions concerning how people communicate with each other. These assumptions are of particular relevance in problems such as the Linda problem, which instead of calculations (as are required, for instance, in Bayesian problems) require semantic inferences about the meaning of the information provided by the experimenter. We believe that these assumptions and inferences reflect intelligent ways in which humans deal with uncertainty.

SOCIAL RATIONALITY

The conversational analysis of the Linda problem is one example of a situation in which adhering to social norms, here conversational maxims, is rational, although it conflicts with classical rationality, defined by many researchers in psychology and economics as adherence to the laws of probability theory and logic. Elsewhere we have challenged this classical vision of human rationality (Chase, Hertwig and Gigerenzer, 1998). The conjunction rule is neither the only rule of internal consistency used as a benchmark of sound reasoning, nor the only one to which people do not seem to measure up. Again, taking into account social rationality — social norms, expectations, and goals — helps us to understand why.

One of the basic principles of internal consistency in choice is known as Property α, which requires that if you choose x over y, you should do so independently of the other alternatives in the choice set. Are violations of Property α irrational? Not necessarily: our social values can sometimes conflict with this principle (Sen, 1993). Consider Property α in the context of social politics at a dinner party. At dessert, it looks as if there are fewer pastries than there are people. By the time the dessert tray gets to you, there is only one pastry left, a chocolate eclair, and you have the choice of taking nothing (x) or taking the eclair (y). If you know that another of the guests has not yet taken a dessert, out of politeness you might choose to have nothing over having the eclair (i.e. x over y). However, if the tray had contained another pastry (z), you might well have chosen to eat that same eclair (y) over having nothing (x). Choosing x from the choice set $\{x, y\}$ and choosing y from the choice set $\{x, y, z\}$ violates Property α. But it is polite, and politeness pays. Not being so could anger others and lessen the chances that they will cooperate with us in the future.

These and other examples of social rationality illustrate why we doubt that psychology can make much progress in understanding people's reasoning by focusing exclusively on syntactical systems such as probability theory and logic. This doubt is as old as experimental psychology itself. As Wilhelm Wundt (1912/1973) expressed it:

> At first it was thought that the surest way would be to take as a foundation for the psychological analysis of the thought-processes the laws of logical thinking, as they had been laid down from the

time of Aristotle by the science of logic ... These norms ... only apply to a small part of the thought processes. Any attempt to explain, out of these norms, thought in the psychological sense of the word can only lead to an entanglement of the real facts in a net of logical reflections. We can in fact say of such attempts, that measured by results they have been absolutely fruitless. They have disregarded the psychical processes themselves ... (pp. 148–149).

ACKNOWLEDGEMENTS

We would like to thank Valerie M. Chase, Daniel G. Goldstein, Denis Hilton, Ulrich Hoffrage, Alejandro López, Peter Sedlmeier, Terry Regier, Anita Todd and Tom Trabasso for many helpful comments.

REFERENCES

Adler, J. 'Abstraction is uncooperative', *Journal for the Theory of Social Behaviour*, **14** (1984), 165–181.
Adler, J. E. 'An optimist's pessimism: Conversation and conjunction', *Posnan Studies in the Philosophy of the Sciences and Humanities*, **21** (1991), 251–282.
Agnoli, F. and Krantz, D. H. 'Suppressing natural heuristics by formal instruction; The case of the conjunction fallacy', *Cognitive Psychology*, **21** (1989), 515–550.
Ahn, W. K. and Bailenson, J. 'Causal attribution as a search for underlying mechanisms: An explanation of the conjunction fallacy and the discounting principle', *Cognitive Psychology*, **31** (1996), 82–123.
Bar-Hillel, M. and Neter, E. 'How alike is it versus how likely is it: A disjunction fallacy in probability judgments', *Journal of Personality and Social Psychology*, **65** (1993), 1119–1131.
Birnbaum, M. H. 'Controversies in psychological measurement', in Wegener, B. (ed.), *Social Attitudes and Psychophysical Measurement* (pp. 401–485), Hillsdale, NJ: Erlbaum, 1982.
Birnbaum, M. H. and Mellers, B. A. 'Bayesian inference: Combining base rates with opinions of sources who vary in credibility', *Journal of Personality and Social Psychology*, **45** (1983), 792–804.
Bless, H., Strack, F. and Schwarz, N. 'The informative functions of research procedures: Bias and the logic of conversation', *European Journal of Social Psychology*, **23** (1993), 149–165.
Budescu, D. V. and Wallsten, T. S. 'Processing linguistic probabilities: General principles and empirical evidence', in Busemeyer, J., Hastie, R. and Medin, D. L. (eds), *Decision Making from the Perspective of Cognitive Psychology* (pp. 275–318), New York: Wiley, 1995.
Cambridge Dictionary of Philosophy, Cambridge: Cambridge University Press, 1995.
Chase, V. M., Hertwig, R. and Gigerenzer, G. 'Visions of rationality', *Trends in Cognitive Sciences*, **2** (1998), 206–214.
Cohen, J. *Statistical Power analysis for the Behavioral Sciences*, 2nd edn, Hillsdale, NJ: Erlbaum, 1988.
Cosmides, L. and Tooby, J. 'Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty', *Cognition*, **58** (1996), 1–73.
Daston, L. J. *Classical Probability in the Enlightenment*, Princeton, NJ: Princeton University Press, 1988.
Donovan, S. and Epstein, S. 'The difficulty of the Linda conjunction problem can be attributed to its simultaneous concrete and unnatural representation, and not to conversational implicature', *Journal of Experimental Social Psychology*, **33** (1997), 1–20.
Duden 'Das große Wörterbuch der deutschen Sprache', Mannheim: Dudenverlag, 1981.
Dulany, D. E. and Hilton, D. J. 'Conversational implicature, conscious representation and the conjunction fallacy', *Social Cognition*, **9** (1991), 85–110.
Erdfelder, E., Bröder, A. and Brandt, M. *Zur Bedeutung des Antwortformats für die Häufigkeit von Konjunktionsfehlern* [The relevance of response format for the frequency of conjunction errors]. Talk presented at the 40th Tagung experimentell arbeitender Psychologen in Marburg, Germany, 1998.
Fabre, J.-M., Caverni, J. P. and Jungermann, H. 'Causality does influence conjunctive probability judgments if context and design allow for it', *Organizational Behavior and Human Decision Processes*, **63** (1995), 1–5.

Fiedler, K. 'The dependence of the conjunction fallacy on subtle linguistic factors', *Psychological Research*, **50** (1988), 123–129.
Gigerenzer, G. 'Why the distinction between single-event probabilities and frequencies is relevant for psychology (and vice versa)', in Wright, G. and Ayton, P. (eds), *Subjective Probability* (pp. 129–161), New York: Wiley, 1994.
Gigerenzer, G. 'On narrow norms and vague heuristics: A reply to Kahneman and Tversky', *Psychological Review*, **103** (1996), 592–596.
Gigerenzer, G. and Hoffrage, U. 'How to improve Bayesian reasoning without instruction: Frequency formats', *Psychological Review*, **102** (1995), 684–704.
Gigerenzer, G., Hoffrage, U. and Kleinbölting, H. 'Probabilistic mental models: A Brunswikian theory of confidence', *Psychological Review*, **98** (1991), 506–528.
Gigerenzer, G. and Murray, D. J. *Cognition as Intuitive Statistics*, Hillsdale, NJ: Erlbaum, 1987.
Gigerenzer, G. and Regier, T. 'How do we tell an association from a rule? Comment on Sloman (1996)', *Psychological Bulletin*, **119** (1996), 23–26.
Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. and Krüger, L. *The Empire of Chance: How probability changed science and everyday life*, Cambridge: Cambridge University Press, 1989.
Gould, S. J. *Bully for Brontosaurus: Reflections in natural history*, New York: Norton, 1992.
Grice, H. P. 'Logic and conversation', in Cole, P. and Morgan, J. L. (eds), *Syntax and Semantics 3: Speech acts* (pp. 41–58), New York: Academic Press, 1975.
Grice, H. P. *Studies in the Way of Words*, Cambridge, MA: Harvard University Press, 1989.
Hacking, I. *The Emergence of Probability*, Cambridge: Cambridge University Press, 1975.
Hertwig, R. *Why Dr Gould's Homunculus doesn't think like Dr Gould: The 'conjunction fallacy' reconsidered*, Doctoral dissertation, Universität Konstanz: Germany. Konstanz: Hartung-Gorre Verlag (1995).
Hertwig, R. and Chase, V. M. 'Many reasons or just one: How response mode affects reasoning in the conjunction problem', *Thinking and Reasoning*, **4** (1998), 319–352.
Hilton, D. J. 'The social context of reasoning: Conversational inference and rational judgment', *Psychological Bulletin*, **118** (1995), 248–271.
Hoffrage, U. and Gigerenzer, G. 'Using natural frequencies to improve diagnostic inferences', *Academic Medicine*, **73** (1998), 538–540.
Jones, S. K., Jones Taylor, K. and Frisch, D. 'Biases of probability assessment: A comparison of frequency and single-case judgments', *Organizational Behavior and Human Decision Processes*, **61** (1995), 109–122.
Juslin, P., Olsson, H. and Winman, A. 'The calibration issue: Theoretical comments on Suantak, Bolger and Ferrell (1996)', *Organizational Behavior and Human Decision Processes*, **73** (1998), 3–26.
Kahneman, D. and Tversky, A. 'On the reality of cognitive illusions', *Psychological Review*, **103** (1996), 582–591.
Kanwisher, N. 'Cognitive heuristics and American security policy', *Journal of Conflict Resolution*, **33** (1989), 652–675.
Koehler, J. J., Gibbs, B. J. and Hogarth, R. M. 'Shattering the illusion of control: Multi-shot versus single-shot gambles', *Journal of Behavioral Decision Making*, **7** (1994), 183–191.
Langer, E. J. 'The illusion of control', *Journal of Personality and Social Psychology*, **32** (1975), 311–328.
Macdonald, R. R. 'Credible conceptions and implausible probabilities', *British Journal of Mathematical and Statistical Psychology*, **39** (1986), 15–27.
Macdonald, R. R. and Gilhooly, K. J. 'More about Linda *or* conjunctions in context', *European Journal of Cognitive Psychology*, **2** (1990), 57–70.
Margolis, H. *Patterns, Thinking, and Cognition: A theory of judgment*, Chicago: University of Chicago Press, 1987.
May, R. S. *Realismus von subjectiven Wahrscheinlichkeiten: Eine kognitionspsychologische Analyse inferentieller Prozesse beim Over-confidence Phänomen*, [Calibration of subjective probabilities: A cognitive analysis of inference processes in overconfidence]. Frankfurt, Germany: Lang, 1987.
Messer, W. S. and Griggs, R. A. 'Another look at Linda', *Bulletin of the Psychonomic Society*, **31** (1993), 193–196.
Morier, D. M. and Borgida, E. 'The conjunction fallacy: A task specific phenomenon?', *Personality and Social Psychology Bulletin*, **10** (1984), 243–252.
Nisbett, R. E., Krantz, D. H., Jepson, C. and Kunda, Z. 'The use of statistical heuristics in everyday inductive reasoning', *Psychological Review*, **90** (1983), 339–363.
Oxford English Dictionary, Oxford: Clarendon Press, 1989.

Politzer, G. and Noveck, I. A. 'Are conjunction rule violations the result of conversational rule violations?', *Journal of Psycholinguistic Research*, **20** (1991), 83–103.
Random House College Thesaurus, New York: Random House, 1984.
Reeves, T. and Lockhart, R. S. 'Distributional versus singular approaches to probability and errors in probabilistic reasoning', *Journal of Experimental Psychology: General*, **122** (1993), 207–226.
Rosch, E. 'Cognitive representations of semantic categories', *Journal of Experimental Psychology: General*, **104** (1975), 192–233.
Sedlmeier, P. 'BasicBayes: A tutor system for simple Bayesian inference', *Behavior Research Methods, Instruments, and Computers*, **29** (1997), 328–336.
Sen, A. 'Internal consistency of choice', *Econometrica*, **61** (1993), 495–521.
Shafir, E. B., Smith, E. E. and Osherson, D. N. 'Typicality and reasoning fallacies', *Memory & Cognition*, **18** (1990), 229–239.
Shapiro, B. J. *Probability and Certainty in Seventeenth-Century England*, Princeton, NJ: Princeton University Press, 1983.
Smith, E. E. and Osherson, D. N. 'Similarity and decision making', in Vosniadou, S. and Ortony, A. (eds), *Similarity and Analogical Reasoning* (pp. 60–75), Cambridge: Cambridge University Press, 1989.
Sniezek, J. A. and Buckley, T. 'Decision errors made by individuals and groups', in Castellan, N. J. (ed.), *Individual and Group Decision Making*, Hillsdale, NJ: Erlbaum, 1993.
Sperber, D. and Wilson, D. *Relevance: Communication and cognition*, 2nd edn, Oxford: Blackwell, 1995.
Stich, S. P. 'Could man be an irrational animal?', *Synthese*, **64** (1985), 115–135.
Teigen, K. H. 'Variants of subjective probabilities: Concepts, norms, and biases', in Wright, G. and Ayton, P. (eds), *Subjective Probability* (pp. 211–238), New York: Wiley, 1994.
Teigen, K. H., Martinussen, M. and Lund, T. 'Linda versus world cup: Conjunctive probabilities in three-event fictional and real-life predictions', *Journal of Behavioral Decision Making*, **9** (1996), 77–93.
Third New International Dictionary, Springfield, MA: Merriam Company, 1967.
Thüring, M. and Jungermann, H. 'The conjunction fallacy: Causality versus event probability', *Journal of Behavioral Decision Making*, **3** (1990), 61–74.
Tversky, A. and Kahneman, D. 'Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment', *Psychological Review*, **90** (1983), 293–315.
Tversky, A. and Koehler, D. J. 'Support theory; A nonextensional representation of subjective probability', *Psychological Review*, **101** (1994), 547–567.
Varey, C. A., Mellers, B. A. and Birnbaum, M. H. 'Judgments of proportions', *Journal of Experimental Psychology: Human Perception and Performance*, **16** (1990), 613–625.
Wundt, W. *An Introduction to Psychology* (translation of *Einführung in die Psychologie* by R. Pinter), New York: Arno, 1912/1973.

Authors' biographies:
Ralph Hertwig is a research scientist at the Max Planck Institute for Human Development, Berlin. He works on ecological rationality, the hindsight bias, and on the comparison of the methodological approaches in psychology and experimental economics.

Gerd Gigerenzer is Director at the Max Planck Institute for Human Development, Berlin. He specializes in ecological rationality, on fast and frugal decision heuristics that make sound inferences, and on the domain-specificity of judgment and decision making.

Authors' address:
Ralph Hertwig and **Gerd Gigerenzer**, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany.

APPENDIX

Probability and frequency versions used by Tversky and Kahneman (1983), Fiedler (1988), and Reeves and Lockhart (1993).

Tversky and Kahneman (1983)

Probability representation
A health survey was conducted in a representative sample of adult males in British Columbia of all ages and occupations. Mr F. was included in the sample. He was selected by chance from the list of participants. Which of the following statements is more probable? (check one)

Mr F. has had one or more heart attacks.
Mr F. has had one or more heart attacks and he is over 55 years old.

Frequency representation
A health survey was conducted in a sample of 100 adult males in British Columbia, of all ages and occupations. Please give your best estimate of the following values:

How many of the 100 participants have had one or more heart attacks?
How many of the 100 participants both are over 55 years old and have had one or more heart attacks?

Fiedler (1988)

Probability representation
Linda is 31 years old, single, outspoken, and very bright ... Please rank order the following statements with respect to their probability.

Linda is a bank teller.
Linda is active in the feminist movement.
Linda is a bank teller and is active in the feminist movement.
[Fiedler presented five other hypotheses in addition to T, F, and T&F]

Frequency representation
Linda is 31 years old, single, outspoken, and very bright ... To how many out of 100 people who are like Linda do the following statements apply?

Linda is a bank teller.
Linda is active in the feminist movement.
Linda is a bank teller and is active in the feminist movement.

Reeves and Lockhart (1993)

'Case-specific' representation
Bill is 34 years old. He is intelligent, but unimaginative, compulsive, and generally lifeless. In school, he was strong in mathematics but weak in social studies and humanities. Rank order the statements by their probabilities:

Bill is an accountant.
Bill plays jazz for a hobby.
Bill is an accountant and plays jazz for a hobby.

'Frequency' representation
Frank, Jake, Adam and Marty live in the same neighborhood but do not know each other. Frank goes to Maxi's Bar and Grill 5 nights a week, Jake goes to Maxi's 3 nights a week, Adam goes to Maxi's 3 nights a week, and Marty goes to Maxi's only 1 night a week. You go to the bar one night. Rank order the statements by their probabilities:

Frank is at the bar.
Marty is at the bar.
Frank is at the bar and Marty is at the bar.
[Reeves and Lockhard presented a total of seven hypotheses.]

