THINKING AND REASONING, 1998, *4* (4), 319–352

Many Reasons or Just One: How Response Mode Affects Reasoning in the Conjunction Problem

Ralph Hertwig

Max Planck Institute for Human Development, Berlin, Germany

Valerie M. Chase

The University of Chicago, USA

Forty years of experimentation on class inclusion and its probabilistic relatives have led to inconsistent results and conclusions about human reasoning. Recent research on the conjunction "fallacy" recapitulates this history. In contrast to previous results, we found that a majority of participants adhere to class inclusion in the classic Linda problem. We outline a theoretical framework that attributes the contradictory results to differences in statistical sophistication and to differences in *response mode*—whether participants are asked for probability *estimates* or *ranks*—and propose two precise cognitive algorithms for ranking probabilities. Our framework allows us to make novel predictions about when and why people adhere to class inclusion. Evidence obtained in several studies supports these predictions and demonstrates that the proposed ranking algorithms can account for about three-quarters of participants' inferences in the Linda problem.

INTRODUCTION

"Our minds are not built (for whatever reason) to work by the rules of probability." The brainteaser that brought Harvard paleontologist Stephen J. Gould (1992, p.469) to this conclusion was the *Linda* problem, in which one reads: "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 antinuclear demonstrations." One is

Requests for reprints should be sent 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

The research reported in this paper was partly supported by a National Science Foundation Graduate Research Fellowship to the second author. We thank Maya Bar-Hillel, Miriam Bassok, Hartmut Blank, Gerd Gigerenzer, Ulrich Hoffrage, Daniel Kahneman, David E. Over, Guy Politzer, Peter Sedlmeier, Peter M. Todd, and Michael Waldmann for many helpful comments.

then asked to rank the following events according to their probability: Linda is a bank teller (B), Linda is active in the feminist movement (F), and Linda is a bank teller and is active in the feminist movement (B&F).

Gould (1992, p.469) humorously describes the conflict: "I know that the third statement [B&F] is least probable, yet a little homunculus in my head continues to jump up and down, shouting at me—'but she can't just be a bank teller; read the description'." Most participants in Tversky and Kahneman's (1983) original study of the Linda problem agreed with Gould's homunculus. Only 10–20% of them ranked B&F as the least probable alternative, while the rest violated the conjunction rule, which states that the mathematical probability of a conjoint event (e.g. B&F) cannot exceed the probability of any of its constituent events. Tversky and Kahneman (1983) called this violation the conjunction "fallacy". The status of the conjunction rule as a norm for single-event probabilities, such as the probability that Linda is a bank teller, has been vigorously debated (see Gigerenzer, 1996; Kahneman & Tversky, 1996), but this normative issue will not concern us here.

In the 15 years since the conjunction fallacy was first demonstrated, many studies have found apparently strong evidence for the conclusion that our minds are not built to work by the rules of probability. Given this evidence, it came as a surprise to us when we found that—contrary to previous results—a majority of a sample of students at the University of Chicago followed the conjunction rule when asked to estimate the event probabilities in the Linda problem. Are the minds of these students, contrary to Gould's conclusion, built to work by the rules of probability? Before addressing this question, we demonstrate that our inconsistent finding is less surprising when viewed in light of 40 years of experimentation on class inclusion and its probabilistic relatives.

A SHORT HISTORY OF CONTRADICTORY FINDINGS AND CONCLUSIONS

Piaget and Inhelder (Inhelder & Piaget, 1959/1969; Piaget, 1952) saw cognitive development as proceeding in an invariant series of stages culminating in logico-mathematical abilities like those of a scientist. One of these abilities is recognising that a set must be larger than any of its subsets (the set-theoretic equivalent of the conjunction rule) to which we refer as reasoning in accord with *class inclusion*. In one experiment, Piaget (1952) showed children a box containing wooden beads, most of which were brown but two white, and asked: "Are there more wooden beads or more brown beads in this box?". Most children under the age of 7 or 8 replied that there were more brown beads. By age 8, however, most responded that there were more wooden beads, in accord with class inclusion. Inhelder and Piaget (1959/1969, p.109 & p.117) concluded that at age 8 and older children "can compare the extension of a part with that of the

whole" and that "this kind of thinking is not peculiar to professional logicians since the children themselves apply it with confidence when they reach the operational level."

Cohen and colleagues (e.g. Cohen & Hansel, 1958), unlike their contemporaries Piaget and Inhelder, focused on the "Achilles heels" (Cohen, Chesnick, & Haran, 1972, p.46) of probability judgement in teenagers and adults. In their view, reasoning in accord with the multiplication rule for independent events, $p(A\&B) = p(A)\,p(B)$, a special case of the conjunction rule, is such an Achilles heel. Cohen and Hansel (1958) had participants estimate conjoint probabilities, for example, the probability of winning two gambles in each of which the player has a 10% chance of winning. Participants' responses were compared with results calculated from the multiplication rule (in this case, .01). Finding that most 12- to 15-year-olds (Cohen et al., 1972; Cohen & Hansel, 1958) overestimated the conjoint probability, Cohen et al. (1972, p.44) concluded that a "grasp of the multiplicative character of a compound probability is far from being in any sense a 'primitive' property of mental processes in relation to the external world."

Working in the 1960s, Peterson and Beach (1967, p.29) argued that the laws of probability theory and statistics could be used to build psychological models that "integrate and account for human performance in a wide range of inferential tasks." They marshalled people's consistency with the multiplication rule as evidence for this argument. In two experiments, Peterson et al. (1965) asked participants to estimate conditional and unconditional probabilities. For example, they presented participants with a trait such as "witty" and asked them to estimate the number out of 100 people who are witty and the number of those witty people who are brave. According to the multiplication rule for dependent events, the products $p(\text{witty})\,p(\text{brave} \mid \text{witty})$ and $p(\text{brave})\,p(\text{witty} \mid \text{brave})$ are equal. Finding correlations of .67 and .90 between the products of such estimates (in their Experiments 1 and 2, respectively), Peterson et al. (1965, p.528) concluded that participants' probability judgements showed "a high degree of internal consistency".

The 1970s and 1980s witnessed the most recent challenge to the view that the laws of reasoning can be modelled by the laws of probability and logic (see Gigerenzer, 1996; Kahneman & Tversky, 1996). Launching a long line of research, Tversky and Kahneman (1974, p.1124) argued that reliance on cognitive heuristics that "reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations" leaves human reasoning prone to "severe and systematic errors". Evidence for this view in the judgement of conjoint probabilities came from studies by Slovic (1969) and Bar-Hillel (1973), who presented problems that demanded extensive application of probability theory (e.g. multiplication of seven event probabilities). Later, a new kind of problem was developed that more closely approximated real-life

situations but was more ambiguous (Hertwig & Gigerenzer, 1997). The most well known problem of this kind is the Linda problem, in which Tversky and Kahneman (1982, 1983) and many other researchers found that only a small minority of people ranked the constituent B as more probable than the conjunction B&F.

REASONING PROBLEMS ARE NOT NEUTRAL MEASUREMENT TOOLS

Why is there such a diversity of results and conclusions across 40 years of experimentation on reasoning in accord with class inclusion, the multiplication rule, and the conjunction rule, to which our results contribute? We argue that this diversity reflects the power of problem structure to direct thought. In so claiming, we build on the work of a number of researchers, including Hogarth (1982, 1987), Payne (1982; Payne, Bettman, & Johnson, 1992), and Lopes (1982), who argued that judgements arise from the interaction between problem structure and cognitive algorithms. Although all designed to test whether people adhere to a set of related rules in logic and probability, the beads problem (Piaget), the gambling problems (Cohen & Hansel, Slovic, and Bar-Hillel), the trait problem (Peterson et al.), and the Linda problem (Tversky & Kahneman) differ not only in content but in various aspects of problem structure, such as *response format* (whether responses are expressed as probabilities or frequencies) and what we refer to as *response mode* (whether participants have to give ranks or estimates). Different problem structures can elicit different cognitive processes, and thereby different judgements—not only on the part of participants, but also on the part of the experimenters who evaluate their performance.

STUDY 1: A PUZZLING RESULT

Teigen, Martinussen, and Lund (1996, p.78) expressed a widely held belief when they concluded that the "conjunction fallacy … [is] very robust and replicable in a number of problem contexts." It is therefore puzzling that we found the usual high percentage of violations of class inclusion in the Linda problem when we asked people to give ranks, but not when we asked them to give estimates. We first describe Study 1, in which we found this surprising result. We then propose a theoretical framework that accounts for how the ranking and estimation response modes elicit different cognitive processes and therefore different judgements. This framework allows us to derive two predictions about when and why people adhere to class inclusion, which we tested in Studies 2 and 3. Finally, we propose two precise algorithms to model reasoning in a ranking version of the Linda problem and evaluate their predictive performance in Study 4.

Method

We tested two groups of participants, each of which received a total of five conjunction problems in one of two fixed orders. For the present purpose, we focus on the problem that all participants received first, namely the Linda problem. Those in the *ranking* group were asked to rank the event probabilities, whereas those in the *estimation* group were asked to estimate them (for precise instructions, see Appendix). The Linda problem included three alternatives: two constituents (B, F) and their conjunction (B&F). As well as being required to judge the probability of each alternative, participants were asked to provide written justifications of their thinking while making the judgements. Order of alternatives was randomised.

Participants. A total of 72 students from the University of Chicago were randomly assigned either to the ranking or to the estimation group (n=36 in each). They were paid volunteers recruited by advertisement from a wide range of disciplines, and were tested in groups of up to five people.

Results

An *inclusion* judgement is defined here as one in which the judged probability of the constituent alternative (e.g. B) is greater than or equal to the probability of the conjoint alternative (B&F). In Study 1, 58% of participants in the estimation group (21 out of 36) gave inclusion judgements (for mean probability estimates, see Table 1). To illustrate how much this result differs from those of previous studies, Fig. 1 plots this percentage alongside those reported in a sample of studies that required participants to rank probabilities in the Linda problem. The percentages of inclusion judgements across 17 conditions in the 10 studies in Fig. 1 range from 5% to 25%. The median is 13%, 45 percentage points lower than the percentage observed in the estimation group in Study 1.

The picture looked different in the ranking group, in which the results of previous studies were replicated. Only 22% of participants (8 out of 36) gave inclusion judgements—a difference of 36 percentage points relative to the estimation group. To estimate the size of the response mode effect, we calculated the phi coefficient and found $\phi = .37$. According to Cohen (1988), this effect is of medium to large size.

SUMMARY OF STUDY 1

In Study 1, we found a context in which inclusion judgements in the Linda problem were much more common than in previous studies. With an estimation response mode, the percentages of inclusion judgements were on average about 45 percentage points higher than those reported in other studies (see Fig. 1). With

TABLE 1
Studies 1, 2, and 3

	Inclusion Judgements		*Violations of Class Inclusion*	
	M	*SD*	*M*	*SD*
Study 1				
B	.16	.18	.14	.09
F	.58	.31	.72	.24
B&F	.09	.12	.38	.20
Study 2				
B	.39	.32	.33	.36
F	.51	.32	.49	.36
B&F	.14	.18	.60	.27
Study 3				
B	.18	.21	.15	.16
F	.56	.30	.66	.25
B&F	.11	.14	.32	.23

Mean probability estimates (M) and standard deviations (SD) for the constituent alternatives B and F and the conjoint alternative B&F in Studies 1, 2, and 3, split by consistency with class inclusion.

a ranking response mode, however, we replicated previous results. How can a difference in response mode account for the increase in inclusion judgements?

Theoretical Framework

We illustrate the present account using the Linda problem, but it applies to all conjunction problems like the Linda problem, such as the Bill problem (Tversky & Kahneman, 1983) and the problems used by Shafir, Smith, and Osherson (1990). Our account rests on three assumptions, which are explicated next.

Inverse Probability Assumption. Previous researchers have proposed that instead of judging the probabilities of the alternatives B, F, and B&F given Linda's characteristics, many people assess the inverse conditional probabilities, for instance, the probability of a person having Linda's characteristics given that the person is bank teller. Explanations for why people assess these inverse conditional probabilities include conceptual pattern recognition (Massaro, 1994), application of the representativeness heuristic (Shafir et al., 1990, Hypothesis 3), and intuitive Bayesian reasoning, in which the hypotheses represent outcomes

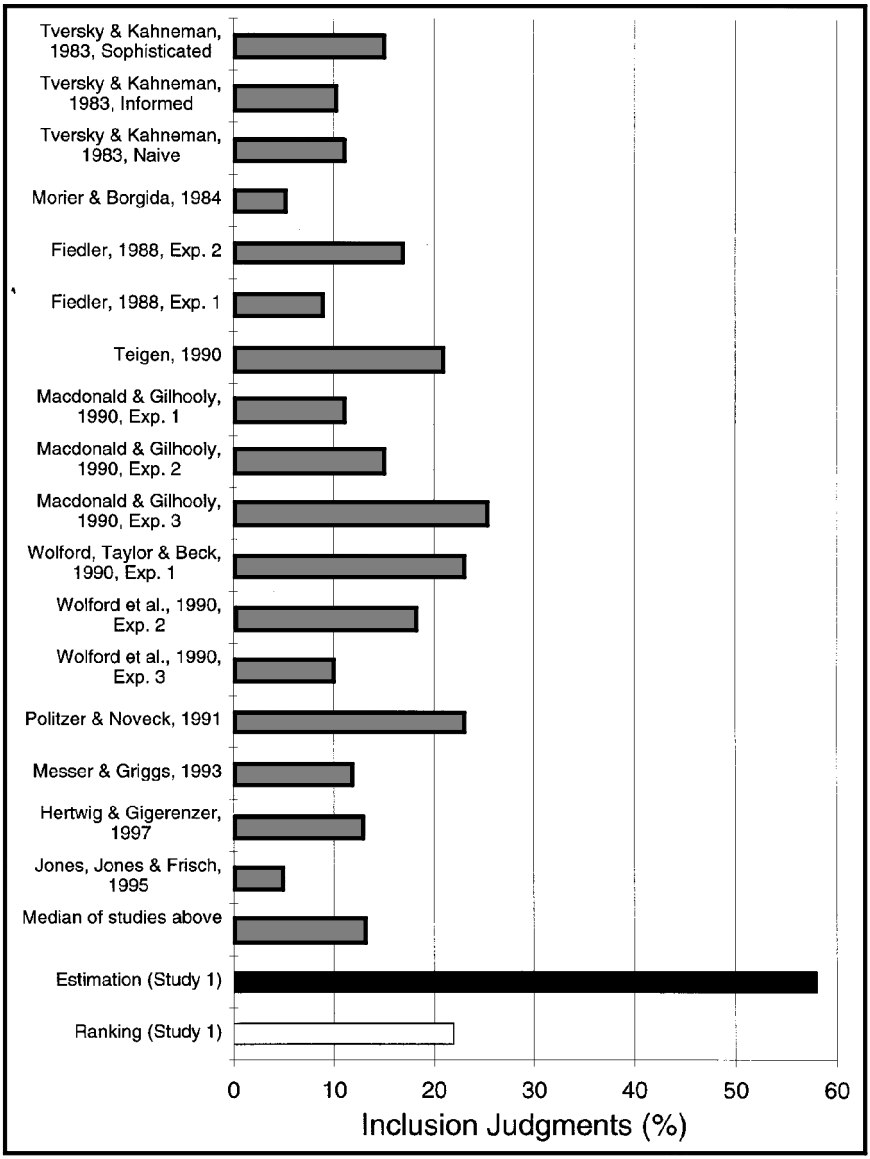

FIG. 1. Percentage of inclusion judgements in the Linda problem across 17 conditions in 10 previous studies (the three results from Tversky & Kahneman, 1983, represent only the first in a series of studies), median percentage in those 17 conditions, and percentages of inclusion judgements in Study 1.

that are assumed to have already occurred (Wolford, Taylor, & Beck, 1990; but see Bar-Hillel, 1991; Fisk, 1996). Although we espouse a different explanation, we too assume that judgements such as those required in the Linda problem are conditioned on the hypotheses rather than on the evidence. We explain why now.

Many natural language terms—including, the term "probability"—are polysemous, that is, have multiple, related meanings. Hertwig and Gigerenzer (1997) argued that people resolve the polysemy of "probability" in the Linda problem by applying norms of social rationality such as the *relevance* maxim (Grice, 1975, 1989). According to the relevance maxim, participants expect the experimenter's "conversational contribution" (i.e. Linda's description) to be relevant. If participants assume this maxim to hold, then they are unlikely to infer that "probability" means mathematical probability (e.g. frequency, percentage, expected value) because such an interpretation would render Linda's description irrelevant to the requested judgement (Adler, 1984, 1991; Hertwig & Gigerenzer, 1997; Hilton, 1995).

Hertwig and Gigerenzer (1997) asked German-speaking participants to paraphrase the term "probability" (which is also polysemous in German) in the Linda problem for an imaginary non-native speaker and found that 82% of their paraphrases were non-mathematical (e.g. possibility, conceivability, credibility). There are several ways of modelling these paraphrases. We propose that most of them can be captured by the notion of *evidential support* (Nozick, 1981). That is, we assume that participants evaluate the degree to which Linda's description provides evidential support for each hypothesis. This assumption implies that judgements in the Linda problem are not systematically constrained by class inclusion because they are conditioned on the hypotheses rather than on the evidence.

Strategy Assumption. The Linda problem can be conceptualised as an inferential task in which three alternatives, B, F, and B&F, are compared to each other on a quantitative dimension, namely the evidential support that Linda's description provides for each alternative. The experimenter provides Linda's values on various dimensions, hereafter referred to as *cues*. For instance, participants are told that Linda's value on the marital status cue is "single". The alternatives' values on these same cues—for example, the probability that feminists are single—are not provided, so that people who consider this information relevant have to retrieve it from memory or to infer it. How might one decide whether Linda provides most support for B, F, or B&F?

We propose two inductive strategies and assume that they are at least partly contingent on response mode: estimation is more likely to activate the *integration* strategy, whereas ranking is more likely to activate the *cue-wise* strategy. In the integration strategy, (a) each alternative (B, F, and B&F) is evaluated independently of the others with respect to evidential support; (b) for each

alternative, Linda's values on the provided cues (marital status, age, and so on) are compared to the memory-retrieved values for the alternative; and (c) the outcomes of these multiple comparisons are expressed as a single value reflecting the evidential support Linda provides for each alternative. In the cue-wise strategy, (a) alternatives are evaluated pair-wise (i.e. B vs. F; B vs. B&F; F vs. B&F) with respect to evidential support; (b) cue values for each alternative in a paired comparison are retrieved from memory and compared to each other cue by cue; and (c) the outcome of these comparisons determines for which alternative Linda provides more evidential support. This assumption—inspired by Tversky's (1969) distinction between additive and additive difference models in the evaluation of multidimensional stimuli—is supported by the consistent finding that response mode can affect the extent to which cognitive processing is alternative-based (as in the integration strategy) or cue-based (as in the cue-wise strategy; e.g. Billings & Scherer, 1988; Rosen & Rosenkoetter, 1976; Schkade & Johnson, 1989; Westenberg & Koele, 1992).

Previous studies on response mode suggest that more cues are searched in estimation than in ranking (e.g. Billings & Scherer, 1988) and that choice (a special case of ranking) is the faster process (e.g. Schkade & Johnson, 1989). We propose a specific variant of the cue-wise strategy based on this evidence and work by Gigerenzer and Goldstein (1996) on *one-reason decision making* in choice problems. Instead of assuming that in the cue-wise strategy the decision maker evaluates a pair of alternatives such as B and F on a cue-by-cue basis and then aggregates (for an example, see Hogarth, 1987), we assume that each choice between alternatives is based exclusively on one reason (i.e. one cue). One-reason decision making uses a minimum of information and eschews complex information integration, yet is sufficient to make the ordinal judgements necessary for a qualitative judgement (choice).

Using the inverse probability and strategy assumptions, we can explain why a cue-wise strategy leads to a high percentage of violations of class inclusion in a ranking version of the Linda problem. If one's interpretation of "probability" can be captured by evidential support, then one can reasonably rank B&F over B because each of Linda's cue values provides more evidence for B&F than for B. This argument, however, should also hold for an estimation version of the Linda problem. How can we explain why a majority of participants in the estimation group in Study 1 judged B&F to be less probable than B?

Rule Assumption. The answer hinges on how B&F is judged. We assume that in the integration strategy, an estimate for B&F is derived by applying *rules* to the constituent estimates (e.g. multiplying p(B) and p(F)). This assumption is based on the following argument. The integration strategy requires that Linda and each alternative be compared on multiple cues specified in Linda's description. To accomplish this, the strategy must retrieve the cue values for each

alternative from memory. In the case of B and F, this should not be problematic because people probably have stored representations of the average bank teller and the average feminist. In the case of B&F, however, the conceptual combination is both novel and incompatible; thus, it is plausible that people do not have a stored representation of it.

As yet, there exists no definitive model of how the mind combines concepts, particularly complex social concepts (for the most precise model of conceptual combination to date, see Smith & Osherson, 1984; Smith, Osherson, Rips, & Keane, 1988). However, it seems fair to suggest that interpretation of conceptual combinations such as "feminist bank teller" involves complex reasoning.[1] Do participants engage in complex reasoning in the Linda problem in order to construct a complete representation of B&F? Gavanski and Roskos-Ewoldsen (1991) argued that they do not. They proposed that participants circumvent having to combine the *concepts* of B and F by combining the *estimates* for B and F using rules instead. Consistent with this argument, they found evidence that violations of class inclusion co-occur with participants' reports of having applied rules that do not follow class inclusion, such as averaging the B and F estimates.

Unlike Gavanski and Roskos-Ewoldsen (1991), we argue that rule application depends on response mode. We propose that in the integration strategy, an estimate for B&F is derived by application of rules to the constituent estimates. Whether or not this process leads to a judgement consistent with class inclusion depends on the particular rule used. It is here that we think that factors such as degree of statistical sophistication play a role by extending the set of available rules to include more complex statistical rules.

There are two reasons to think that rules are unlikely to be applied to judging B&F in a ranking response mode. First, as long as participants can infer some cue values (or as we assume, only one cue value) of B&F that discriminate between alternatives—for example, that feminist bank tellers tend to be more outspoken than bank tellers—they have a basis on which to decide B&F's rank relative to B and F. Thus, the cue-wise strategy does not require the construction of a detailed representation of B&F: cue values that are impossible to retrieve from memory or difficult to infer can be left out of the comparisons between alternatives. Second,

[1]It is widely accepted that most combined concepts cannot be characterised as the intersection of the properties of their constituents (Chater, Lyon, & Myers, 1990; Hampton, 1988; Osherson & Smith, 1981). Alternative characterisations have been suggested. Murphy and Medin (1985, p.306), for instance, proposed that the interpretation of a conceptual combination "may be thought of as a hypothesis generated by background theories". Such a hypothesis may be particularly complex in the case of social concepts, for which assignment of a cue value such as "outgoing" requires consideration not only of a person's behaviour but of the situation (Smith, 1988). There is also evidence that when reasoning about combinations of social concepts, people infer emergent cue values (i.e. values not part of either constituent concept; Hastie, Schroeder, & Weber, 1990) and causal accounts (Kunda, Miller, & Claire, 1990).

applying rules in ranking would be fairly complex.[2] For these reasons, we assume that the cue-wise strategy does not involve rule application.

Summary. The proposed theoretical framework rests on three assumptions. First, participants' probability judgements are conditioned on the hypotheses rather than on the evidence. (We model these judgements using evidential support.) Second, there are two strategies for making inferences in conjunction problems whose elicitation is at least partly contingent on response mode. Specifically, the estimation mode is assumed to trigger an integration strategy and the ranking mode to trigger a cue-wise strategy. Third, the integration strategy involves rule application, which circumvents combination of the constituent concepts. Rule application can lead to systematic adherence to class inclusion, depending on the rules used. In contrast, the cue-wise strategy does not involve rule application and in the Linda problem leads to violations of class inclusion.

Predictions

The proposed theoretical framework allows us to derive two predictions:

Prediction 1: Mode-dependent Adherence to Class Inclusion. In a between-subjects design, participants will be more likely to give inclusion judgements when asked to estimate probability than when asked to rank probability.

Statistical sophistication may interact with response mode in influencing the likelihood of adherence to class inclusion. We assume that statistical education extends the set of combination rules by adding computationally complex rules for judging the conjoint alternative (e.g. the multiplication rules for dependent and independent events) to a person's repertoire. Whether the greater rule repertoire of sophisticated participants actually leads to more inclusion judgements depends on whether statistically naive participants use rules that are simple yet also conform to class inclusion. If statistically naive participants apply such rules, then the effect of response mode on adherence to class inclusion will be independent of degree of statistical sophistication.

Prediction 2: Mode-dependent Rule Use. Participants will be more likely to apply rules to judge the probability of the conjoint alternative in estimation than in ranking. This prediction follows from the assumption that the estimation response mode activates an integration strategy that triggers rule application, whereas the ranking response mode does not.

[2]To calculate B&F's rank, one would have to combine the ranks of B and F and then update the ranks accordingly. For instance, if one ranks F and B "1" and "2", respectively, and applies the averaging rule, then B&F is temporarily ranked "1.5". Next, one must update the ranks to reflect the new order: "1" for B, "2" for B&F, and "3" for F.

STUDY 2: THE EFFECTS OF RESPONSE MODE AND STATISTICAL SOPHISTICATION

In Study 2, we test Prediction 1. If Study 2 supports Prediction 1, we will have replicated the finding in Study 1 that percentage of inclusion judgements varies with response mode. To find out whether and how response mode interacts with statistical sophistication, in Study 2 we assembled participants who we expected would have different degrees of statistical education: some were university students and others were passers-by on a busy street (henceforth *laypeople*). Participants' statistical background was measured by their performance on three textbook probability problems (henceforth *background* problems) that required application of the multiplication rule for independent events (for background problems, see Appendix).

Method

All participants first received the Linda problem. Of the 100 laypeople, 50 were asked to rank and 50 to estimate probability. Of the 152 students, 73 were asked to rank and 79 to estimate probability. After completing the Linda problem, participants received the three background problems, in which they were required to describe the steps that would lead to an estimate but not to do the calculations. After finishing the background problems, the laypeople were asked for their age and whether or not they had ever attended a class in which probability theory was taught. The Linda problem and the ranking and estimation instructions were the same as those used in Study 1 (see Appendix). Order of alternatives in the Linda problem was randomised; order of the three background problems was kept constant.

Participants. Of the 252 participants, 152 were students at the Universities of Munich and Leipzig and 100 were passers-by in downtown Munich. The participants in both Munich groups were compensated for their participation (with candy), and were individually tested. Participants at the University of Leipzig received course credit for their participation, and were tested in groups of up to five people.

Results

Statistical Sophistication. As expected, the students were more statistically sophisticated than the laypeople as measured by performance on the three background problems. Only 22% of the laypeople (22 out of 100) gave responses consistent with the multiplication rule (which for brevity's sake we call *correct* solutions) for at least one of these problems (only 13, 6, and 3 participants correctly solved one, two, and three problems, respectively). In contrast, 76% of the students (116 out of 152) correctly solved at least one problem (37, 44, and 35

correctly solved one, two, and three problems, respectively). This result is not surprising given that most of the laypeople had finished school long before (mean age: 45 years), and only 20% could remember ever having attended a class in which probability theory was taught.

Inclusion Judgements. As measured by performance on the background problems, the participants in Study 2 varied widely in statistical sophistication. We therefore analysed the percentage of inclusion judgements as a function of both response mode and statistical sophistication. With respect to statistical sophistication we distinguished between two groups of participants: those who did not solve a single background problem correctly (the *naive* group) and those who solved at least one background problem correctly (the *sophisticated* group). Only 17% of naive participants in the ranking group (9 out of 53) gave inclusion judgements, whereas 46% in the estimation group (28 out of 61) did so, a difference of 29 percentage points ($\phi = .31$; a medium effect size). We found a similar effect of response mode among sophisticated participants: whereas 40% in the ranking group (28 out of 70) gave inclusion judgements, 65% in the estimation group (44 out of 68) did so, a difference of 25 percentage points ($\phi = .25$; a small to medium effect size). In Prediction 1, we predicted that a larger percentage of participants in the estimation group would give inclusion judgements than participants in the ranking group. In terms of both percentage differences and effect sizes, the results of Study 2 support this prediction, and replicate the results obtained in Study 1 (for mean estimates, see Table 1).

Among the sophisticated participants, does the effect of response mode depend on the number of background problems solved correctly? Figure 2 shows the percentage of inclusion judgements as a function of the number of correct solutions. For those who correctly solved one, two, and three problems, respectively, the effects of response mode on percentage of inclusion judgements are 29 ($\phi = .28$), 16 ($\phi = .16$), and 33 ($\phi = .32$) percentage points. Except for the relatively small effect for participants who correctly solved two problems, the results indicate that the effect of response mode is fairly stable over a range of statistical sophistication.

In concluding that the effect of response mode is independent of statistical sophistication, we are not claiming that sophistication has no effect on reasoning in accord with class inclusion. It does in two ways. First, the percentages of inclusion judgements are 23 ($\phi = .25$) and 19 ($\phi = .19$) percentage points higher for the sophisticated than the naive participants in the ranking and estimation groups, respectively. Second, statistical sophistication affected the complexity of rules used to judge the conjoint alternative. To infer rule use in the estimation group, one can take each participant's constituent estimates—i.e. $p(\mathrm{B})$ and $p(\mathrm{T})$—and calculate her conjoint estimate as if she had applied each of a number of rules that yield single-point predictions (e.g. the multiplication rule rather than

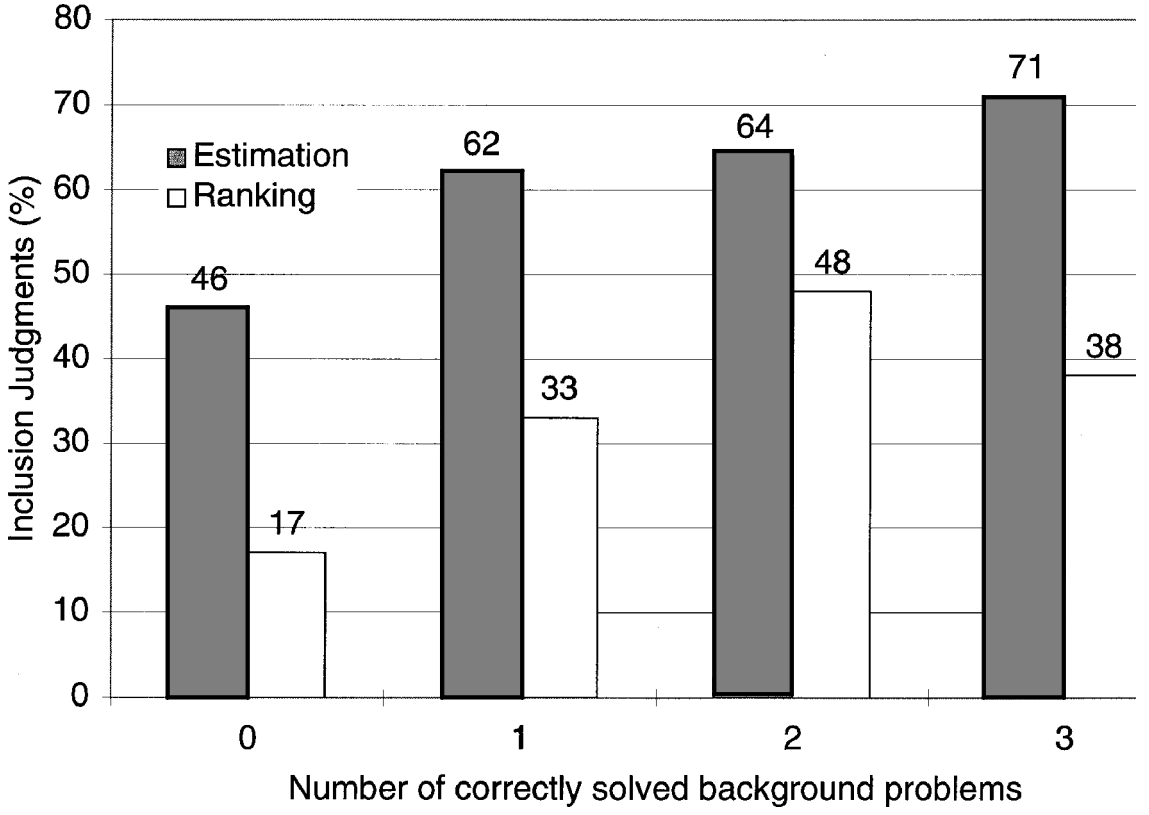

FIG. 2. Percentage of inclusion judgements as a function of response mode and number of background problems solved correctly.

a weighted averaging rule). If the estimate predicted by a particular rule is consistent with the actual conjoint estimate, then one can consider this as evidence that the rule was applied. For those participants in the estimation group who gave inclusion judgements (n = 72), we checked whether the conjoint estimates conformed to the multiplication rule for independent events—i.e. $p(B\&F) = p(B)\,p(F)$—or to the *ceiling* rule, by which the conjoint estimate is set equal to the lower of the two constituent estimates—i.e. $p(B\&F) = Min(p(B), p(F)$. Although both rules lead to inclusion judgements, the multiplication rule is computationally more complex than the ceiling rule. Therefore, when both rules correctly predicted a participant's conjoint estimate—e.g. $p(F) = 1$, $p(B) = 0$, and $p(B\&F) = 0$—we inferred that the ceiling rather than the multiplication rule had been applied.

About one-third of sophisticated participants who conformed to class inclusion gave conjoint estimates consistent with the ceiling rule (36%; 16 out of 44), while about one-quarter gave estimates consistent with the multiplication rule (23%; 10 out of 44). More than half of naive participants who conformed to class inclusion gave conjoint estimates consistent with the ceiling rule (57%, 16 out of 28), but none seems to have applied the multiplication rule. Finally, a minority of participants in both groups adhered to class inclusion without applying either rule. These participants (14% in the sophisticated group and 18% in the naive group) simply gave a conjoint estimate of 0 although both of their constituent estimates were greater than 0.

Summary

The results of Study 2 support Prediction 1. Participants are more likely to give inclusion judgements in probability estimates than ranks. We obtained this finding in several different populations: an American student population (at the University of Chicago), a German student population (at the Universities of Munich and Leipzig), and a German lay population (on the streets of Munich). We also found that the effect of response mode is independent of statistical sophistication as measured by performance on three background problems requiring application of the multiplication rule for independent events. A larger percentage of both sophisticated and naive participants gave inclusion judgements when estimating than ranking probabilities in the Linda problem.

However, statistical sophistication affected the overall percentage of inclusion judgements. Sophisticated participants gave more inclusion judgements than naive participants across response modes (a 21 percentage point difference). Sophistication also affected rule complexity in the estimation group. Whereas about one-quarter of sophisticated participants appeared to apply the multiplication rule for independent events, no naive participant did so, suggesting that statistical education is needed for this rule to be in one's repertoire. The ceiling rule, in contrast, was used by an appreciable percentage of participants in both

groups. We propose that the estimation response mode and statistical sophistication combine to increase the percentage of inclusion judgements in conjunction problems such as Linda.

STUDY 3: IS RULE USE CONTINGENT ON RESPONSE MODE?

We now explore the issue of rule use more thoroughly. In Study 3, we tested whether people will be more likely to apply rules to judging the probability of the conjoint alternative in estimation than in ranking, as stated in Prediction 2. In addition, we tested whether or not inclusion judgements in estimation co-occur with evidence that rules consistent with class inclusion were applied to judging the conjoint probability.

Method

Each participant was randomly assigned to one of two groups, one of which was instructed to rank probability and the other to estimate probability. The same version of the Linda problem was used in both groups as in Study 1, with B, F, and B&F presented in a randomised order. Participants in the estimation group were asked to give on-line written justifications of their judgements, and those in the ranking group to fill out a postexperiment questionnaire that asked them to report rule use.

We administered the postexperiment questionnaire in the ranking and not the estimation group because of the differential informativeness of ranks and estimates. To infer rule use from estimates, one can take each participant's constituent estimates and predict the conjoint estimate assuming each of a number of rules that yield single-point predictions was applied, as we did in Study 2. Thus, the estimates and written justifications in the estimation group can serve as indicators of rule use.

Ranks, in contrast, are not informative with respect to rule use because a particular set of ranks (e.g. F>B&F>B) can be predicted by multiple rules. Therefore, we explicitly asked participants in the ranking group about rule use. The postexperiment questionnaire explained that one of many ways to derive ranks is to estimate p(B), p(F), and p(B&F), and then to rank the estimated probabilities. To find out whether participants did this, we asked them if they had used any of the following rules: averaging (mean of constituents), addition (sum of constituents), subtraction (difference between constituents), multiplication (product of constituents), averaging and adjusting (up or down), and ceiling (match between conjoint estimate and lower of the two constituents). Finally, we included a catch-all category that allowed participants to report whether they used other rules (e.g. averaging ranks). Note that the ceiling and multiplication

rules necessarily lead to inclusion judgements, whereas the other rules either necessarily lead to or at least can lead to violations. Participants in the ranking group were asked to indicate the rule that they used (if any).

Participants. Each of the 136 participants was randomly assigned to one of two groups (n = 77 in estimation and n = 59 in ranking). They were paid volunteers from the University of Chicago recruited by advertisement, and were tested in groups of up to five people.

Results

Before reporting the results bearing on Prediction 2, we report a final test of Prediction 1. In Study 3, 58% of participants in the estimation group (45 out of 77) gave inclusion judgements, whereas only 24% in the ranking group (14 out of 59) did so—a difference of 34 percentage points (ϕ = .35; for mean estimates, see Table 1). Combined with the results obtained in Studies 1 and 2, this constitutes strong evidence that inclusion judgements are more likely in an estimation response mode than a ranking response mode.

Mode-dependent Rule Use. As a first step, we inferred that a participant had applied a rule only if (a) the written justification (estimation group) or post-experiment questionnaire (ranking group) specified that a rule was used, and (b) the given estimates or ranks were consistent with the rule reported. This analysis stacks the deck against our prediction that rule use will be more common in the estimation group than in the ranking group, because participants in the latter group were explicitly questioned about rule use. By this stringent criterion, 36% of participants in the estimation group (28 out of 77) showed evidence of using a rule, 8 of them reporting the ceiling rule and 20 the multiplication rule. In the ranking group, 20% (12 out of 59) showed evidence of rule use, 7 of them reporting the multiplication rule and 5 reporting either averaging or averaging and adjusting up or down.

As a second step, we checked each conjoint judgement in the estimation group to see whether it was consistent with any of the rules included in the ranking group's postexperiment questionnaire that yield single-point estimates (e.g. averaging). If a conjoint estimate satisfied the criteria for more than one rule, it was counted as an instance of whichever rule assumes the least amount of computation. By this criterion, another one-tenth of participants in the estimation group (10%; 8 out of 77) gave conjoint estimates consistent with the ceiling rule, while a few more gave estimates consistent with addition (2 out of 77) and averaging (1 out of 77).

Aggregating across these two measures of rule use, 51% of participants in the estimation group (39 out of 77) and only 20% of participants in the ranking group

(12 out of 59) seem to have applied rules to derive their judgements for the conjoint alternative. By Cohen's (1988) classification, the contingency of rule use on response mode is of medium effect size ($\phi = .31$). The effect is smaller if one considers only the written justifications in the estimation group, but is in the predicted direction ($\phi = .17$). The results of Study 3 thus support Prediction 2.

Summary

Study 3 tested whether rules are more often applied to probability estimates than to probability ranks (Prediction 2). The results are consistent with this prediction. Roughly half of participants in the estimation group either spontaneously reported using rules or gave conjoint estimates consistent with a rule, whereas only one-fifth of participants in the ranking group reported rule use when directly asked.

About one-quarter of participants in the estimation group in Study 3 (26%; 20 out of 77) reported using the multiplication rule, a percentage that is much closer to the 23% found in the sophisticated group in Study 2 (where rule use was inferred from the constituent estimates) than to the 0% found in the naive group in Study 2. We argue that the participants in Study 3 look like the sophisticated group in Study 2 because students at the University of Chicago are relatively statistically sophisticated. According to *Barron's Profiles of American Colleges* (1994), 85% of first-year undergraduate students at the University of Chicago in 1993 (some of whom actually participated in Study 3) scored over 600 (centred) on the mathematics section of the Scholastic Aptitude Test. In addition, the undergraduate curriculum requires each student to take at least two courses in the mathematical sciences. In a background questionnaire completed by a subgroup of the University of Chicago students in Studies 1 and 3, 78% of participants (74 out of 95) reported having taken at least one course in one of these areas, with an average of 2.3 courses per participant. In short, there is evidence that this population has statistical sophistication, which might explain why about one-quarter of the estimation group in Study 3 reported using the multiplication rule.

Study 3 contributes to a growing body of evidence that rules are sometimes applied to making probability judgements in conjunction problems. Gavanski and Roskos-Ewoldsen (1991) measured rule use using a postexperiment questionnaire that included a wider range of rules than found in Study 3. Averaged across problems, 75% and 71% of participants in their studies (in their Experiments 1 and 2, respectively) reported using a rule to make their conjoint estimates. In addition, Chase and Bassok (1998) performed a rule analysis of probability estimates in the Linda and other conjunction problems in two studies and found that more than half of conjoint estimates conformed to a rule, a result comparable to that found in the estimation group in Study 3.

In the next section, we propose and test two algorithms that model the cognitive processes of the cue-wise strategy, which according to our framework is triggered by the ranking response mode.

STUDY 4: TWO REALISATIONS OF THE CUE-WISE STRATEGY

In the cue-wise strategy, the judge is assumed to evaluate a pair of alternatives (such as B and F) on the basis of one cue. As observed earlier, the cue-wise strategy is therefore an example of one-reason decision making (Gigerenzer & Goldstein, 1996). Before proposing two precise realisations of it, we define a good reason in the Linda problem as one that maximises the difference in evidential support between alternatives.

Evidence has been defined in many ways (for a review, see Schum, 1994). The definition advanced by Nozick (1981, p.252) has the virtues of simplicity and generality. Nozick proposed the difference between likelihoods $s = p(E \mid H) - p(E \mid \neg H)$ as a measure of the force that evidence E provides hypothesis H as opposed to ¬H. A positive s indicates the degree to which E supports H, and a negative s indicates the degree to which E supports ¬H. Applied to the Linda problem, this measure captures how much support each alternative (e.g. F) receives from a piece of evidence E (e.g. Linda is outspoken). For instance, if $p(\text{outspoken} \mid F) = .5$ and $p(\text{outspoken} \mid \neg F) = .2$, then $s_F = .3$ indicates the degree to which Linda's outspokenness supports the hypothesis "Linda is a feminist". If $p(\text{outspoken} \mid B\&F) = .3$ and $p(\text{outspoken} \mid \neg(B\&F)) = .2$, then $s_{B\&F} = .1$ indicates the degree to which Linda's outspokenness supports the hypothesis "Linda is a bank teller and is active in the feminist movement". Now one can compare the support for F (.3) and B&F (.1), which should result in choosing F as more probable than B&F (i.e. $s_F - s_{F\&B} > 0$). This calculation, however, can be simplified in the Linda problem in the following way.

Consider the second term of the support measure, $p(E \mid \neg H)$. In the Linda problem, ¬H can refer to, for instance, all "non-bank tellers" (¬B) or all "non-feminist non-bank tellers" (¬(B&F)). As these hypotheses include most of the population (e.g. the vast majority of people are not bank tellers), the probabilities of all of these catch-all hypotheses conditioned on the evidence are essentially the same. Suppose we draw a random sample of 1000 people from the general population. Five of the people are bank tellers, and one of these five is a feminist bank teller and is outspoken, whereas only one of the four non-feminist bank tellers is outspoken. If the base rate of outspokenness in the sample of 1000 people is 300, then $p(\text{outspoken} \mid B)$ equals .4 (i.e. 2/5) and $p(\text{outspoken} \mid \neg B)$ equals .3 (i.e. 298/995), whereas $p(\text{outspoken} \mid B\&F)$ equals 1 (i.e. 1/1) and $p(\text{outspoken} \mid \neg(B\&F))$ equals .3 (i.e. 299/999). That is, spoken $p(\text{outspoken} \mid \neg(B\&F))$ both approximate the base rate of outspokenness, and the equation:

$$[p(\text{outspoken} \mid \text{B}) - p(\text{outspoken} \mid \text{–B})] - [p(\text{outspoken} \mid \text{B\&F}) - p(\text{outspoken} \mid \neg(\text{B\&F}))]$$

reduces to:

$$p(\text{outspoken} \mid \text{B}) - p(\text{outspoken} \mid \text{B\&F}).$$

Therefore, the choice between B and B&F can be made quite simply. If p(outspoken | B) =.4 and p(outspoken | B&F) = 1, then, $s_{B,\,B\&F} = -.6$, and B&F should be chosen as more probable than B. Note that the probability p(E | B&F) can be greater than, less than, or equal to p(E | B), and thus need not follow class inclusion. Using this evidential support measure, we now propose the first cue-wise algorithm.

The One-reason Algorithm

The policy of the *One-reason* algorithm is "Take the best cue and ignore the rest". It assumes that the cues provided in the Linda problem are tested for their ability to discriminate between B, F, and B&F. The choice between each pair of alternatives x and y is based on the "best" cue; that is, the cue that maximises the difference in evidential support between alternatives (i.e. $s_{x,\,y}$). Because in the Linda problem the cues that can be used to make a decision are not generated by the decision maker but are instead presented by the experimenter, we do not assume that the cues are *a priori* ranked according to $s_{x,y}$ for each pair of alternatives. The One-reason algorithm consists of the following steps:

1. *Set up the first choice*. Choose at random the first pair of alternatives to be compared.
2. *Test for the best cue*. Test which cue maximises the absolute value of $s_{x,y}$.
3. *Rule for choice*. Choose the alternative to which the best cue points; if no cue discriminates, then choose between the two alternatives at random.
4. *Set up the second choice*. Compare the alternative chosen from the first pair with the remaining alternative.
5. *Repeat Steps 2 and 3*.
6. *Determine probability ranks*. If the initially chosen alternative "wins" again, then rank it "1" and compare the two "losers" to determine their relative ranks (i.e. repeat Steps 2 and 3). If the initially chosen alternative loses, then rank it "2" and the other alternatives "1" and "3", respectively.

In the One-reason algorithm, there is no integration across cues, alternatives are not evaluated independently, and choices between pairs of alternatives are determined solely by the best cue.

The One-reason algorithm is *non-compensatory* in that no combination of other cues can outweigh the best cue. There is experimental evidence that people use non-compensatory strategies in judgement tasks. For instance, Hayes-Roth and Hayes-Roth (1977) found that classification performance is best predicted by the "property-set" model, which classifies solely on the basis of the most diagnostic property.[3] Russo and Dosher (1983) argued that cue-wise processing requires less cognitive effort because it involves procedures that simplify the decision, and reported evidence that choices are based on a single most discriminating dimension. Finally, Billings and Scherer (1988) found that participants looked up less information in a choice condition in which they had to select the best candidate for a job than in a judgement condition in which they had to evaluate each candidate on a 7-point scale.

The Minimalist Algorithm

One might object that the assumption of the One-reason algorithm that all seven cues in the Linda problem are tested in the search for the best cue is unrealistic. A less computationally demanding alternative to the One-reason algorithm is the *Minimalist* algorithm, which simply picks a cue at random and chooses between the two alternatives based on the evidential support the cue provides each alternative. It differs from the One-reason algorithm only in Steps 2 and 3, which we reformulate as Steps 2′ and 3′ below:

2′. *Random selection*. For the two alternatives, select a cue from Linda's description at random and compute the evidential support it provides.
3′. *Rule for choice*. Choose the alternative to which the cue points; if the cue does not discriminate, then select another cue at random.

Although the Minimalist and One-reason algorithms are only two possible realisations of a cue-wise strategy, we focus on them because they contrast most sharply with the integration strategy by relying on just one cue (either the best cue or a randomly selected one). In Study 4, we test the extent to which each algorithm can account for participants' ranks in the Linda problem. Study 4 was administered to participants in the ranking group of Study 3, who after giving ranks were asked to provide the information that the algorithms require as input.

---

[3]The diagnosticity of a property set (e.g. red wagon) for any class was conceptualised as an increasing function of its associative strength to that class (e.g. "what Jane likes") and a decreasing function of its associative strength to other classes (e.g. "what Sue likes"). The property-set model assumes that the strength of association between a property set and a class is expressed as a likelihood estimate.

The algorithms' predictions are made for each *individual* participant in Study 4 based on his or her estimation of the parameters.

Method

Participants were given a booklet in which they were instructed to estimate the frequencies of positive values for each cue specified in Linda's description in the three alternative classes B, F, and B&F (see Appendix). For instance, they were asked: "Think of all *female bank tellers* in the United States. Imagine a representative sample of 100 of them. How many of these 100 female bank tellers do you expect to be *31 years old*?" This information was needed to calculate evidential support cue by cue. Participants responded to 21 such questions, one for each of the seven cue values specified in Linda's description (i.e. 31 years old, single, outspoken, very bright, philosophy major, deeply concerned with discrimination and social justice, participated in antinuclear demonstrations) in each of the three alternative classes (B, F, and B&F). To control for order effects, we used all six possible orders of the three alternatives and two fixed orders of the seven cue questions within each alternative (the order shown in the Appendix, and its reverse).

Participants. Participants were the 59 participants in the ranking group of Study 3.

Results

We focus on the two comparisons—B vs. B&F and F vs. B&F—in which class inclusion can be violated. Figure 3 shows the percentages of these actual judgements that the One-reason algorithm correctly predicted. It performed well in predicting the violations in the B vs. B&F choice (39 out of 41) and the inclusion judgements in the F vs. B&F choice (41 out of 52). However, it predicted few of the inclusion judgements in the B vs. B&F choice (4 out of 18), and few of the violations in the F vs. B&F choice (2 out of 7). Averaged across the two comparisons that could lead to violations, the One-reason algorithm predicted about three-quarters (73%) of participants' actual ranks.

To test the predictions of the Minimalist algorithm, we chose one of the seven cues in the Linda problem at random for each participant; if the first cue did not discriminate between alternatives, we chose another at random until one that discriminated was found. We then recorded the alternative to which the discriminating cue gave greater evidential support. We repeated this procedure 10 times for each participant and for each pair-wise comparison. Averaged across all trials and all participants, the average number of accurate predictions made by the Minimalist algorithm is 70% (see Fig. 3). The performance of the Minimalist algorithm barely suffered from assuming that the choice is based on a randomly

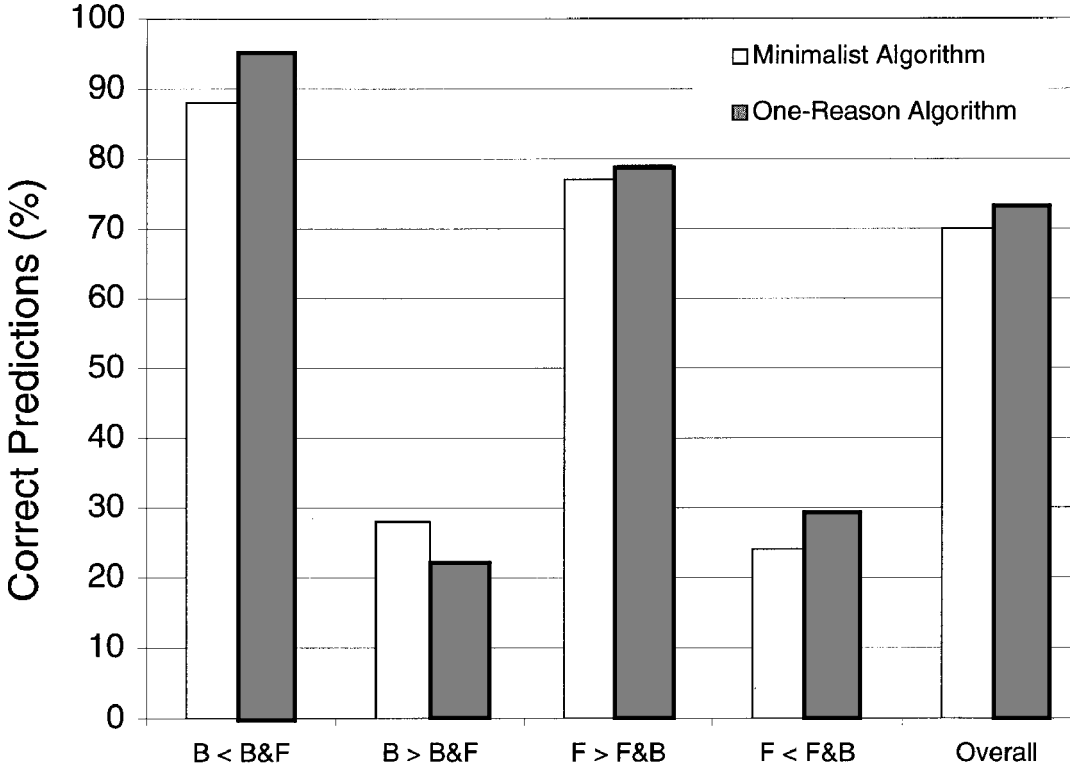

FIG. 3. Percentage of correct predictions of the following types of actual judgement by the One-reason and Minimalist algorithms: B<B&F, B>B&F, F>B&F, F<B&F.

selected cue instead of assuming, as the One-reason algorithm does, that all cues are tested in the search for the best cue (70% vs. 73% correct predictions).

Why is the performance of the two algorithms so similar? We suggest that cue redundancy, which is pronounced in the Linda problem, allowed the Minimalist algorithm to approximate the performance of the One-reason algorithm closely: averaged across participants and pair-wise comparisons, nearly five (4.8) out of seven cues pointed to the same alternative. Thus, in the Linda problem, it hardly matters whether an algorithm determines its judgement on the basis of the best cue or a randomly selected cue.

Summary

The One-reason and Minimalist algorithms were able to account for about three-quarters of participants' ranks. This performance is not perfect. In particular, the algorithms predicted only a few of the inclusion judgements in the B vs. B&F choice. We can think of at least one possible reason. The results of Study 3 show that a few participants reported having applied the ceiling rule in the ranking response mode. Because the proposed algorithms do not include a step involving rule application, they cannot predict ranks derived from correct rules. We can nevertheless aim to have them capture most people's responses, which they do.

GENERAL DISCUSSION

We now discuss the implications of the results reported here, as well as some of the assumptions involved in our explanatory framework.

The Conjunction Fallacy: Impervious to Education?

According to Piatelli-Palmarini (1994, p.140), cognitive "illusions" such as the conjunction fallacy are inevitable, one reason being that they are "independent of intelligence and education". We argue that the findings reported in research on reasoning in accord with class inclusion do not justify such a categorical statement. Although in their early work on the conjunction fallacy Tversky and Kahneman (1982, p.93) concluded that statistical sophistication has only a "negligible effect", based on their subsequent studies they later concluded that "recognition of the decisive nature of rules distinguishes … different levels of statistical sophistication" (Tversky & Kahneman, 1983, p.300). Moreover, other researchers have found a non-negligible effect of statistical sophistication in conjunction problems.

One of Tversky and Kahneman's (1983) later studies was performed on social science graduate students at the University of California, Berkeley, and Stanford University who had taken several statistics courses. Participants were asked to judge the probability of B and B&F in the Linda problem on a 1–9 rating scale.

An unusually high percentage of them—64%—gave inclusion judgements. To explain this result, Tversky and Kahneman (1983) invoked two factors they suggested that can highlight the inclusion relation and thereby increase the percentage of inclusion judgements: (1) *directness*, that is, whether each participant judges both the conjoint and constituent alternatives in a within-subject design (direct test) or different participants judge the critical alternatives in a between-subjects design (indirect test); and (2) *transparency* , that is, whether the critical alternatives are presented alone (direct transparent test) or in a longer list of alternatives (direct subtle test). They concluded that statistically sophisticated participants such as those in the study just described particularly profit from directness and transparency of the inclusion relation.

Agnoli and Krantz (1989) and Fisk and Pidgeon (1997) found that training in statistical principles can make people reason in accord with class inclusion. Using Venn diagrams, they taught participants about set relations such as inclusion and disjunction. Agnoli and Krantz (1989, Experiment 1) found that the percentage of inclusion judgements increased from 44% to 73% with training (averaged across subgroups in their Table 3), and Fisk and Pidgeon (1997, Table 2) found a smaller increase from 37% to 47% (for "likely–unlikely" conjunctions such as B&F). Because they gave one group of trained participants an indirect test of the conjunction problem, Agnoli and Krantz (1989) were also able to address the question of whether statistical training only benefits participants in direct tests (as implied by Tversky and Kahneman's explanation). Although smaller than in the direct test, Agnoli and Krantz (1989) found a beneficial effect of training even in the indirect test.

These researchers manipulated statistical sophistication in the laboratory by training some participants and not others. Benassi and Knoth (1993), in contrast, assessed participants' level of sophistication by asking them to solve problems in probability theory (see also Donovan & Epstein, 1997). They gave participants three test problems that required application of the multiplication rule for independent events (e.g. judging the probability of getting a head on the first flip and a head on the second flip of a fair coin) to assess their sophistication, and then gave them a direct, transparent conjunction problem similar to the Linda problem (the Dan problem). They found that only 21% of participants who did not answer any of the test problems in accord with the multiplication rule gave ranks that obeyed class inclusion in the Dan problem, whereas 45% of those who answered all three test problems according to the multiplication rule gave inclusion judgements. (For studies in which researchers assessed participants' sophistication simply by their field and course level, as did Tversky & Kahneman, 1983, see Wolford et al., 1990, and Reeves & Lockhart, 1993.)

In Study 2, we found an effect of statistical sophistication on percentage of inclusion judgements that was of small to medium size. Taken together with the

studies just cited, this constitutes evidence that statistical sophistication plays a role in reasoning in accord with class inclusion. Thus, contrary to Piatelli-Palmarini's (1994) claim, the conjunction "fallacy" is not independent of statistical education. We know of only one explanation of the effect of statistical sophistication, namely Tversky and Kahneman's suggestion that it allows people to benefit from directness and transparency. Alternatively, statistical education might have the side effect of providing people with a technical language in which the term "probability" is restricted to its mathematical meanings. Lacking such education, people are free to infer any of the term's acceptable meanings in natural language, many of which are non-mathematical (Hertwig & Gigerenzer, 1997).

Response Mode Effects in Conjunction Problems

There is evidence beyond our studies that response mode matters to inclusion judgements. We know of two studies that directly compared an estimation and a ranking response mode in the Linda problem and found an effect. Morier and Borgida (1984) observed an increase of 15 percentage points in inclusion judgements when participants were asked to give probability estimates compared to ranks. Similarly, Hertwig and Gigerenzer (1997) observed an increase of 20 percentage points. In addition, we can tentatively compare the results of the many studies that employed only a ranking response mode to the few in which only an estimation response mode was used (e.g. Fisk, 1996, Study 1; Fisk & Pidgeon, 1997; Wells, 1985). Comparing the percentage of inclusion judgements in these studies (40% in the first two studies, see Fisk, 1996, Table 3; 27% in Wells, 1985) to the median percentage reported in the studies displayed in Fig. 1 (13%), one again finds that people are more likely to give inclusion judgements when asked to estimate than to rank probabilities. The only study of which we know in which no effect of response mode was found is that of Reeves and Lockhart (1993), who tested problems other than Linda.

In Studies 1–3, the difference between the percentage of inclusion judgements observed in probability estimates and ranks was consistently substantial. We do not know why Reeves and Lockhart (1993) did not find any effect of response mode, or why others (e.g. Morier & Borgida, 1984) found only a small effect. However, we know of at least one condition under which the effect of response mode is likely to be overestimated. In Study 2, 17% of naive participants gave inclusion judgements in ranks, whereas 65% of sophisticated participants (i.e. those who solved at least one background problem correctly) did so in estimates. We suggest that this 48 percentage point difference reflects the combined effects of response mode and statistical sophistication. Although such a combination of effects cannot explain the discrepancy between our findings and those of Reeves and Lockhart (1993), it could explain why one finds different effect sizes across studies depending on whether one or both variables are held constant.

The Generality of the Support Measure

We used evidential support as an index of the degree to which the given evidence supports one hypothesis over another. Specifically, we borrowed Nozick's (1981) difference measure, $s = p(E \mid H) - p(E \mid \neg H)$. Researchers have found evidence that people engage in such likelihood subtraction in several reasoning contexts, for instance, in solving Bayesian inference problems (Gigerenzer & Hoffrage, 1995). Moreover, Nozick's (1981) measure is closely related to measures proposed to underlie information selection in hypothesis testing and classification decisions.

How can we quantify the diagnosticity of a question (e.g. "Have you ever visited Graceland?"), that is, its informativeness with respect to the rejection or acceptance of a hypothesis (e.g. that the respondent is an Elvis groupie)? As a measure of a question's diagnosticity Slowiaczek, Klayman, Sherman, and Skov (1992) proposed the following simple estimate: Given two initially equiprobable hypotheses, diagnosticity is proportional to the simple difference between the probability of a "yes" (or a "no") answer under H and under ¬H. This difference amounts to Nozick's (1981) support measure, and as Slowiaczek et al. (1992) pointed out, also correlates highly with the expected log likelihood measure (Klayman & Ha, 1987).

Skowronski and Carlston (1987) suggested that another judgement process—social impression formation—is a probabilistic categorisation in which people use cues (evidence) to assign a person to one or more trait categories (hypotheses). As a measure of the diagnosticity of cues (e.g. "reports all taxable income") for the traits of interest (e.g. honesty), they proposed a likelihood ratio in which the numerator is the probability that a person with trait H will exhibit a behaviour E, and the denominator is the sum of this probability and that of an actor with the opposite trait ¬H performing the same behaviour $p(E \mid H) / ((p(E \mid H) + p(E \mid \neg H))$. We found that the performance of the One-reason and Minimalist algorithms barely changes if one uses Skowronski and Carlston's (1987) measure instead of Nozick's (1981).

Evidential support and cue diagnosticity rest on the notion of *category validity*, which is defined as the conditional probability that an entity has some cue value given that it belongs to a class. It is the converse of *cue validity*, the probability that an entity is a member of a class given that it has that cue value (Corter & Gluck, 1992). There are at least two reasons to believe that these measures should be based on category rather than cue validity when applied in contexts such as the Linda problem. First, Medin, Wattenmaker, and Michalski (1987) showed that people tend to emphasise category over cue validity. Second, cue validity necessarily increases for more inclusive classes (Murphy, 1982), and thus—unlike category validity—cannot account for violations of class inclusion.

Representativeness versus One-reason Algorithm and Evidential Support

How do the One-reason and Minimalist algorithms, which assume that judgements in the Linda problem are based on evidential support, differ from the representativeness explanation? Although in the Linda problem they are difficult to distinguish empirically, we see differences as well as common features.

The most important difference is that the One-reason and Minimalist algorithms rely exclusively on one cue, whereas the representativeness heuristic uses all of the given information. Both the representativeness heuristic (Tversky & Kahneman, 1983) and precise models of it proposed to apply in the Linda problem (Shafir et al., 1990; Smith & Osherson, 1989) are integration rather than cue-wise strategies in that they assume that the similarity between Linda and the representation of each alternative is calculated *across* the cues provided by the experimenter. Shafir et al. (1990, p.237) specified one realisation of representativeness thus: "To assess, for example, the typicality of Linda in the category *bank teller*, the subject computes two weighted sums, namely: (1) the weighted sum of the features common to Linda's description and the category *bank teller*, and (2) the weighted sum of the features found in one of the latter two feature sets but not the other. These two weighted sums are then combined by a linear rule." Based on Tversky's (1977) contrast rule, Smith and Osherson (1989) proposed a precise version of this featural computation process in the Linda problem.

We did not use Smith & Osherson's (1989) model of the representativeness heuristic to account for the probability estimates in our studies, which according to our framework should be modelled with an integration strategy, for three reasons. First, the representativeness explanation of reasoning in the Linda problem is based on the assumption that people construct a novel conceptual combination such as feminist bank teller and apply a feature comparison strategy to B&F just as they do to the constituent alternatives B and F. We demonstrated, however, that many people in our studies applied rules to estimating the conjoint alternative. Second, the prototypes of B and B&F presumed by Smith and Osherson (1989) in applying their model to the Linda problem lead it to predict a low percentage of inclusion judgements such as those found in the previous studies displayed in Fig. 1. In our studies, however, it is a *high* percentage of inclusion judgements in estimation that requires explanation. According to our framework, it is rule application rather than the integration strategy *per se* that leads to the high percentage of inclusion judgements in estimation. Of course, we could still have used the representativeness heuristic to model people's constituent estimates, but here we stumble on our third reason for not applying the representativeness model: it sacrifices simplicity for exhaustiveness. In it, (a) each prototype (e.g. bank teller) contains slots for a variety of attributes (e.g. education), each of which is specified by three parameters (diagnosticity, values,

and votes), (b) the comparison between prototype and instance (e.g. Linda) follows complex integration rules, and (c) four free parameters (e.g. the weights of the common and distinctive features) are available to fit the data. We suppose that this is why no one has yet conducted an empirical test of this model in the conjunction problem, including Smith and Osherson (1989) themselves, who assumed parameter values instead of measuring them.

The One-reason and Minimalist algorithms have more in common with an alternative version of representativeness outlined (and apparently favoured) by Shafir et al. (1990): "Subjects take the typicality of instance *i* in category *C* to be the judged probability of something's being *i*-like." According to Shafir et al. (1990, p.238), this hypothesis "portrays the subject's poor probability judgment as the result of calculating the wrong probability." Applied to the Linda problem, it implies that people estimate p(Linda | B) rather than p(B | Linda). In other words, this version of representativeness shares the inverse probability assumption integral to our theoretical framework. However, the two accounts attribute the fact that people assess this inverse probability to different sources. Whereas this version (Shafir et al., 1990, p.238) of the representativeness explanation seems to attribute it to the fact that "intuitive probability estimates are not extensional", we believe that it stems from the fact that people infer nonmathematical meanings of the polysemous term "probability" in the Linda problem (based on reasoning guided by conversational maxims).

Although there are seven cues available in the Linda problem, we cannot distinguish the predictions of the One-reason algorithm from those of representativeness because of the high redundancy across cues (e.g. most of the cues point to F rather then B). To resolve their predictions, one could construct problems in which several cues point to alternative *a*, and just one cue—the best one (i.e. that which maximises evidential support)—points to alternative *b*. In this case, an integration strategy such as representativeness predicts that people will choose *a* (assuming that the sum of all cues outweighs the best cue) whereas the noncompensatory One-reason algorithm predicts that people will choose *b*.

Are Our Minds Built to Work by the Rules of Probability?

As we and others before us have demonstrated, mind design interacts with problem design. The structure of problems shapes the cognitive processes people use to solve them, and the multitude of problem designs therefore leads to a multitude of contradictory conclusions about our ability to reason according to probability theory and logic. The Linda problem illustrates this state of affairs. Drawing on research to date, we can design a Linda problem that reliably elicits reasoning that violates class inclusion. Components of this design include use of the polysemous term "probability" (Hertwig & Gigerenzer, 1997), a ranking response mode, and an invitation to participants to infer that B means B&¬F

(Adler, 1991; Dulany & Hilton, 1991; Hertwig, 1995). We can also construct a Linda problem that as reliably elicits reasoning consistent with class inclusion. Such a design would replace the polysemous term "probability" with the term "frequency" (Fiedler, 1988; Hertwig & Gigerenzer, 1997; Tversky & Kahneman, 1983), provide an estimation response mode, and block inferences such as B&¬F (Politzer & Noveck, 1991; Tversky & Kahneman, 1983).

The lesson we draw is that by choosing to realise a problem using a particular design we wittingly or unwittingly adopt a particular perspective on the mind. We should remember that whatever perspective we take allows us to observe only some of the reasoning processes in which people can engage.

Manuscript received 28 February 1997
Revised manuscript received 5 March 1998

REFERENCES

Adler, J.E. (1984). Abstraction is uncooperative. *Journal for the Theory of Social Behaviour*, *14*, 165–181.

Adler, J.E. (1991). An optimist's pessimism: Conversation and conjunction. *Poznan Studies in the Philosophy of the Sciences and Humanities*, *21*, 251–282.

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

Bar-Hillel, M. (1973). On the subjective probability of compound events. *Organizational Behavior and Human Performance*, *9*, 396–406.

Bar-Hillel, M. (1991). Commentary on Wolford, Taylor, and Beck: The conjunction fallacy. *Memory & Cognition*, *19*, 412–414.

Barron's Educational Series, Inc. (1994). *Barron's Profiles of American Colleges* (20th Edn.). New York: Barron's Educational Series, Inc.

Benassi, V.A., & Knoth, R.L. (1993). The intractable conjunction fallacy: Statistical sophistication, instructional set, and training. *Journal of Social Behavior and Personality*, *8*, 83–96.

Billings, R.S., & Scherer, L.L. (1988). The effect of response mode and importance on decision-making strategies: Judgment versus choice. *Organizational Behavior and Human Decision Processes*, *41*, 1–19.

Chase, V. M., & Bassok, M. (1998). *Rule-based reasoning in similarity and probability judgment.* Manuscript in preparation.

Chater, N., Lyon, K., & Myers, T. (1990). Why are conjunctive categories overextended? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 497–508.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Edn.). Hillsdale, NJ: Erlbaum.

Cohen, J., Chesnick, E.I., & Haran, D. (1972). A confirmation of the inertial-Ψ effect in sequential choice and decision. *British Journal of Psychology*, *63*, 41–46.

Cohen, J., & Hansel, C.E.M. (1958). The nature of decisions in gambling: Equivalence of single and compound subjective probabilities. *Acta Psychologica*, *13*, 357–370.

Corter, J.E., & Gluck, M.A. (1992). Explaining basic categories: Feature predictability and information. *Psychological Bulletin*, *111*, 291–303.

Donovan, S., & Epstein, S. (1997). 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*, 1–20.

Dulany, D.E., & Hilton, D.J. (1991). Conversational implicature, conscious representation, and the conjunction fallacy. *Social Cognition*, *9*, 85–110.

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

Fisk, J.E. (1996). The conjunction effect: Fallacy or Bayesian inference? *Organizational Behavior and Human Decision Processes*, *67*, 76–90.

Fisk, J.E., & Pidgeon, N. (1997). The conjunction fallacy: The case for the existence of competing heuristic strategies. *British Journal of Psychology*, *88*, 1–27.

Gavanski, I., & Roskos-Ewoldsen, D.R. (1991). Representativeness and conjoint probability. *Journal of Personality and Social Psychology*, *61*, 181–194.

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

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

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

Gould, S.J. (1992). *Bully for brontosaurus: Further reflections in natural history*. Penguin Books.

Grice, H.P. (1975). Logic and conversation. In P. Cole & J.L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). New York: Academic Press.

Grice, H.P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Hampton, J.A. (1988). Overextension of conjunctive concepts: Evidence for a unitary model of concept typicality and class inclusion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 12–32.

Hastie, R., Schroeder, C., & Weber, R. (1990). Creating complex social conjunction categories from simple categories. *Bulletin of the Psychonomic Society*, *28*, 242–247.

Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning and Verbal Behavior*, *16*, 321–338.

Hertwig, R. (1995). *Why Dr. Gould's homunculus doesn't think like Dr. Gould: The "conjunction fallacy" reconsidered*. Doctoral dissertation, Universität Konstanz, Germany. Konstanz: Hartung-Gorre Verlag.

Hertwig, R., & Gigerenzer, G. (1997). *The "conjunction fallacy" revisited: How intelligent inferences look like reasoning errors*. Manuscript submitted for publication.

Hilton, D.J. (1995). The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin*, *118*, 248–271.

Hogarth, R.M. (Ed.). (1982). *New directions for methodology of social and behavioral science: Question framing and response consistency*. San Francisco: Jossey-Bass.

Hogarth, R.M. (1987). *Judgement and choice: The psychology of decision* (2nd Edn.). London: Wiley.

Inhelder, B., & Piaget, J. (1969). *The early growth of logic in the child*. New York: Norton. [Original work published in 1959, *La genèse des structures logiques elementaire*.]

Jones, S.K., Jones, K.T., & Frisch, D. (1995). Biases of probability assessment: A comparison of frequency and single-case judgments. *Organizational Behavior and Human Decision Processes*, *61*, 109–122.

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

Klayman, J., & Ha., Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*, 211–228.

Kunda, Z., Miller, D.T., & Claire, T. (1990). Combining social concepts: The role of causal reasoning. *Cognitive Science*, *14*, 551–577.

Lopes, L.L. (1982). *Toward a procedural theory of judgment*. University of Wisconsin, Department of Psychology, Wisconsin Information Processing Program.

Macdonald, R.R., & Gilhooly, K.J. (1990). More about Linda: or Conjunctions in context. *European Journal of Cognitive Psychology*, *2*, 57–70.

Massaro, D.W. (1994). A pattern recognition account of decision making. *Memory & Cognition*, *22*, 616–627.

Medin, D.L., Wattenmaker, W.D., & Michalski, R.S. (1987). Constraints and preferences in inductive learning: An experimental study of human and machine performance. *Cognitive Science*, *11*, 299–339.

Messer, W.S., & Griggs, R.A. (1993). Another look at Linda. *Bulletin of the Psychonomic Society*, *31*, 193–196.

Morier, D.M., & Borgida, E. (1984). The conjunction fallacy: A task specific phenomenon? *Personality and Social Psychology Bulletin*, *10*, 243–252.

Murphy, G.L. (1982). Cue validity and levels of categorization. *Psychological Bulletin*, *91*, 174–177.

Murphy, G.L., & Medin, D.L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289–316.

Nozick, R. (1981). *Philosophical explanations*. Oxford: Clarendon Press.

Osherson, D.N., & Smith, E.E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, *9*, 35–58.

Payne, J.W. (1982). Contingent decision behavior. *Psychological Bulletin*, *92*, 382–402.

Payne, J.W., Bettman, J.R., & Johnson, E.J. (1992). Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology*, *43*, 87–131.

Peterson, C.R., & Beach, L.R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*, 29–46.

Peterson, C.R., Ulehla, Z.J., Miller, A.J., Bourne, L.E., & Stilson, D.W. (1965). Internal consistency of subjective probabilities. *Journal of Experimental Psychology*, *70*, 526–533.

Piaget, J. (1952). *The child's conception of number*. New York: Norton.

Piattelli-Palmarini, M. (1994). *Inevitable illusions: How mistakes of reason rule our minds*. New York: John Wiley & Sons.

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

Reeves, T., & Lockhart, R.S. (1993). Distributional versus singular approaches to probability and errors in probabilistic reasoning. *Journal of Experimental Psychology: General*, *122*, 207–226.

Rosen, L.D., & Rosenkoetter, P. (1976). An eye fixation analysis of choice and judgment with multiattribute stimuli. *Memory and Cognition*, *41*, 747–752.

Russo, J.E., & Dosher, B.A. (1983). Strategies for multiattribute binary choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 676–696.

Schkade, D.A., & Johnson, E.J. (1989). Cognitive processes in preference reversals. *Organizational Behavior and Human Decision Processes*, *44*, 203–231.

Schum, D.A. (1994). *The evidential foundations of probabilistic reasoning*. New York: Wiley.

Shafir, E.B., Smith, E.E., & Osherson, D.N. (1990). Typicality and reasoning fallacies. *Memory & Cognition*, *18*, 229–239.

Skowronski, J.J., & Carlston, D.E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology*, *52*, 689–699.

Slovic, P. (1969). Manipulating the attractiveness of a gamble without changing its expected value. *Journal of Experimental Psychology*, *79*, 139–145.

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

Smith, E.E. (1988). Concepts and thought. In R.J. Sternberg & E.E. Smith (Eds.), *The psychology of human thought* (pp.19–49). Cambridge: Cambridge University Press.

Smith, E.E., & Osherson, D.N. (1984). Conceptual combination with prototype concepts. *Cognitive Science*, *8*, 337–361.

Smith, E.E., & Osherson, D.N. (1989). Similarity and decision making. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp.60–75). Cambridge: Cambridge University Press.

Smith, E.E., Osherson, D.N., Rips, L.J., & Keane, M. (1988). Combining prototypes: A selective modification model. *Cognitive Science*, *12*, 485–527.

Teigen, K.H. (1990). To be convincing or to be right: A question of preciseness. In K.J. Gilhooly, M.T.G. Keane, R.H. Logie, & G. Erdos (Eds.), Lines of thinking (Vol. 1). Chicester, UK: Wiley.

Teigen, K.H., Martinussen, M., & Lund, T. (1996). Linda versus world cup: Conjunctive probabilities in three-event fictional and real-life predictions. *Journal of Behavioral Decision Making*, *9*, 77–93.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*, 31–48.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp.84–98). Cambridge: Cambridge University Press.

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

Wells, G.L. (1985). The conjunction error and the representativeness heuristic. *Social Cognition*, *3*, 266–279.

Westenberg, M.R.M., & Koele, P. (1992). Response modes, decision processes and decision outcomes. *Acta Psychologica*, *80*, 169–184.

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

APPENDIX

Materials Used in Study 1

Estimation Instruction. Your task will be to estimate the numerical probability of each statement. Express your probability estimate in terms of a number in the range 0 to 1, where 0 means minimal probability and 1 means maximal probability. You are free to use the whole range (including 0 and 1); both decimal estimates (e.g., .10) and fractional estimates (e.g. 1/10) are acceptable.

Ranking Instruction. Your task will be to rank the three statements that follow each person's description according to their probabilities. Assign a rank of "1" to the statement you think is most probable, "2" to the second most probable statement, and a rank of "3" to the least probable statement.

Linda Problem. 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.

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

Background Problems Used in Study 2

Coin Problem. Imagine you throw a fair coin two times. What is the probability that it will come up "heads" two times in a row?

Die Problem. Imagine you throw a fair die three times. What is the probability that it will come up "6" all three times?

Diabetic Problem. What is the probability of being both a diabetic and a smoker, if 1 in 100 people in the general population is a diabetic and 3 in 10 people are smokers (assume that the two events are independent)?

Materials Used in Study 4

[Only the request for frequency estimates for B alternative shown here.]

Think of all the *female bank tellers* in the U.S. Take a representative sample of 100 of them. We will ask you some questions about this representative sample. Please write down your best guesses.

How many of these 100 female bank tellers do you expect
- as a student participated in anti-nuclear demonstrations? _____out of 100
- as a student were deeply concerned with issues of discrimination and social justice? _____out of 100

How many of these 100 female bank tellers do you expect to have
- majored in philosophy? _____out of 100

How many of these 100 female bank tellers do you expect to be
- bright? _____out of 100
- outspoken? _____out of 100
- single? _____out of 100
- 31 years old? _____out of 100