

The Bayesian logic of frequency-based conjunction fallacies

Momme von Sydow¹

Department of Psychology, University of Göttingen, Gosslerstr. 14, 37073 Göttingen, Germany

ARTICLE INFO

Article history:

Received 30 October 2008

Received in revised form

9 December 2010

Available online 2 March 2011

Keywords:

Probability judgments

Conjunction fallacy

Inductive logic

Paradoxes of predication

Bayesian models

Rationality

ABSTRACT

An inductive, pattern-sensitive Bayesian logic (BL) is proposed as a normative and descriptive model for probability judgments about hypotheses involving probabilistic logical connectives. The model explains a specific class of frequency-based conjunction fallacies (CFs). It is suggested that the pattern probabilities calculated by BL may serve as a criterion of noisy-logical predication, resolving some paradoxes of predication. The model is developed for frequency information in 2×2 contingency tables. According to standard probability theory, a violation of the conjunction rule, $P(A) \geq P(A \wedge B)$ (e.g., $P(\text{ravens are black}) \geq P(\text{ravens are black AND they can fly})$), is always a fallacy. A frequentist interpretation of probability has exculpated participants from committing CFs when one is concerned with single events. Here a pattern-based Bayesian interpretation of probabilities of (noisy) dyadic logical predications is elaborated, predicting frequency-based but rational 'CFs'. BL formalizes the probabilities of logical patterns, integrating over noise levels. BL, for instance, predicts double CFs, differential sample-size effects, and pattern sensitivity. Three experiments provide a first corroboration that BL is also an adequate empirical model to predict logical probability judgments based on 2×2 contingency tables. BL may shed light on the more general rationality debate.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

Since Aristotle, the ability to reason in accordance with the laws of the crystalline machinery of formal logic has been central to the idea that humans are rational animals. Nonetheless, systematic deviations from the norms of logic have been found, which has led researchers to suggest a probabilistic approach to human reasoning (e.g., Oaksford & Chater, 2007). Clearly, uncertainty is a ubiquitous and ineradicable aspect of empirical knowledge, whether one is concerned with diagnosing diseases, making legal judgments, or estimating the underlying characteristics of a person in order to predict future behavior. Intuitions about probability, however, have also been shown to deviate radically from the norms of the standard probability calculus (e.g., Gilovich, Griffin, & Kahneman, 2002; Nickerson, 2004). This article is concerned with fundamental problems encountered at the intersection of logical judgment and probability estimation. How should probabilities be assigned to nested logical hypotheses (problem of inclusion), so

that they can serve as criteria for valid logical predication? How can one account for exceptions in logical prediction (problem of exceptions)? In particular, the article posits a rational basis for frequency-based conjunction fallacies, when people aim to access overall situations in an uncertain world.

The organization of this article is as follows. The first section begins by introducing the type of task investigated, that is, judging the most likely logical rule from a set of instances in a 2×2 contingency table. The second section gives an outline of problems or paradoxes arising from frequency-based probabilistic criteria of adequate predication. It is argued that only pattern probabilities of noisy-logical rules may provide the needed adequacy criterion for logical predication to resolve these problems. It is further suggested that a particular commonsense notion of probability may roughly correspond to such pattern probabilities of noisy rules. In the following section, a technical model called *Bayesian logic* is presented, formalizing pattern probabilities of logical connectives integrating over noise levels. This model should apply to situations concerning overall characterization rather than the specification of a particular extension or subset. Three experiments are then presented, to assess several predictions resulting from the Bayesian logic model, using full frequency information. In the General Discussion, the main theories of the conjunction fallacy are addressed, which primarily have been concerned only with single-event probabilities. It is argued that all major theories of the conjunction fallacy cannot account for the results, even if they may in fact be applied to frequency-based contingency-table tasks.

E-mail addresses: momme.von-sydow@uni-heidelberg.de, Momme@von-Sydow.de.

¹ Momme von Sydow is now at the Department of Psychology, University of Heidelberg, Germany. Portions of the presented results or related results have been presented previously at the 29th, 30th, and 31st Annual Conferences of the Cognitive Science Society (August, 2007, Nashville; July, 2008, Washington; August, 2009, Amsterdam) and at the 41st and 42nd Annual Society for Mathematical Psychology Conferences (July, 2008, Washington; August, 2009, Amsterdam).

Table 1

Logical truth tables of three binary logical connectives (conjunction, affirmation, and inclusive disjunction).

A	B	$A \wedge B$ A and B	A A (whether B or not B)	$A \vee B$ A or B
T	T	T	T	T
T	F	F	T	T
F	T	F	F	T
F	F	F	F	F

Table 2

Observed absolute frequencies x in the four cells a, b, c, d of a contingency table summarizing the co-variation between two binary events.

	B	Non-B
A	x_a	x_b
Non-A	x_c	x_d

Finally, the results are related to the larger debate on fallacious probability judgments.

2. Contingencies and the probability of general logical hypotheses

This article deals with a seemingly simple kind of frequency-based probability-judgment task concerning general logical predications based on data summarized in 2×2 contingency tables.

Predication attributes a predicate to a subject. Here we concentrate on general predication (e.g., “ravens are black”) concerned with a class X of entities, rather than particular predication (e.g., “this raven, ‘Hugin,’ is black”). We focus on the logical relationship between two given predicates A and B of the same class X (e.g., “are black” and “can fly”). Sentences such as “ravens are black and they can fly” involve the use of the common linguistic conjunction “and”, given the class X in question. Such sentences are often taken to correspond to a logical conjunction ($A \circ B \mid X = A \wedge B \mid X$; the symbol “ \circ ” here refers to an unspecified logical connective). Likewise, other ordinary language terms are often taken to represent connectives of logic: “or” (\vee), “either or” ($> <$), “if then” (\rightarrow), and so forth (see Table 1).

The tasks address the probability of alternative logical predications that may be nested: the conjunction is for instance included by the affirmation (see the problem of inclusion). The tasks are concerned with contingent predications based on induction rather than on deduction alone (formal predications). Participants obtain frequency information about the co-occurrence of two attributes A and B . In the simple tasks, participants are shown the data in a 2×2 contingency table (Table 2). Participants judge the probability of particular hypotheses (rating tasks) or choose the most probable hypothesis (choice tasks).

A contingency table may, of course, summarize trial-by-trial data as well. The transparent contingency table format chosen here, however, allows for the disentangling of effects of contingency judgment from effects of memory or subjective representations of frequencies that are intensively studied in contingency learning tasks (De Houwer & Beckers, 2002; Kruschke, 2008; Pinenò & Miller, 2007). Furthermore, the task is concerned neither with contingency assessment in general nor with the strength of correlations (Hattori & Oaksford, 2007; Kao & Wassermann, 1993; McKenzie & Mikkelsen, 2007; White, 2002) or causal links (Cheng, 1997; Hagemayer, Sloman, Lagnado, & Waldmann, 2007; Oberauer, Weidenfeld, & Fischer, 2007; Waldmann, 2007). The task seems in fact to be most closely related to the intense debate on probability judgments concerning logically nested hypotheses and so-called “conjunction fallacies” (e.g., Crupi, Fitelson, & Tentori, 2008; Fisk & Slattery, 2005; Gigerenzer, 1994, 1996; Hertwig, Benz, & Krauss, 2008; Hintikka,

2004; Kahneman & Frederick, 2005; Lagnado & Shanks, 2002; Neace, Michaud, Bolling, Deer, & Zecevic, 2008; Nilsson, 2008; Sides, Osherson, Bonini, & Viale, 2002; Sloman, Over, Slovak, & Stibel, 2003; Tversky & Kahneman, 1983; Wedell & Moro, 2008). This paper in fact expounds and tests a model of a specific kind of conjunction fallacy based on frequency information in a 2×2 contingency table. Nevertheless, since the conjunction fallacy debate has mainly dealt with single-event probabilities, it will be only addressed in the General Discussion.

3. Paradoxes of predication and the basic idea

How should probability judgments be understood in tasks such as the ones above? An initial assumption is that the term “probability” is polysemous (Hertwig & Gigerenzer, 1999; Sloman & Over, 2003; Teigen, 1994). That is, different goals may require different formalizations of probability, even if they are all basically consistent with the axioms of probability (Kolmogorov, 1933).

According to the predominant extensional and frequentist understanding – here signified by an index (P_E) – probability is defined as the proportion of confirmative cases (the extension of a set) relative to all observed cases in a universe of discourse X . Accordingly, in our example, the probability of the sentence “ravens are black and they can fly” is the relative frequency of the event “black & fly” given all previously observed ravens: $P_E(A \text{ and } B \mid X) = f(A \text{ and } B) / N$, with N being the frequency of all events in X (all data in this universe of discourse). Extensional probability is certainly a valid normative account – if one is interested in the relative frequency of a particular subset alone. Moreover, at first glance, extensional probability appears coherent with the common epistemological and communicative goal of logical predication – i.e., the relaying of valid information about a given situation. Correspondingly, it has been argued recently that the traditional logical truth criterion of predication (advocated by philosophers such as Frege, Russell, Whitehead, Wittgenstein, and Popper) may need to be replaced by a high-probability criterion in order to allow for exceptions (Schurz, 2001, 2005; cf. Adams, 1986). Illustrated formulaically, the predictability of the above sentence about ravens required that the ratio of events corroborating that ravens are black and that they can fly is above a given threshold: $P_E(A \text{ and } B \mid X) > \varphi$, with φ above 0.5 (normally this is restricted to only a necessary and not a sufficient criterion).

Although the approach to probabilities developed here shares the goal of retaining a rational criterion for predication based on probability, it improves on the extensional realization of this idea. That is, I propose a technical formalization at a “meta”-level of the probability of patterns of probability (P_P). It is posited that this refers to an important but largely neglected class of intensional human probability judgments. The proposed model is modeled based on the assumed communicative goal of providing an overall valid and informative logical description of underlying propensities, at best in a single sentence. The results of this formalization are more completely called *logical pattern probabilities* (here, $P_P(A \circ B \mid D)$); and their aim is to provide the probability of underlying noisy-logical explanatory propensities, given the data.

In what follows, then, extensional probabilities (P_E) and (logical) pattern probabilities (P_P) are considered in the context of establishing criteria for adequate predication. It is explained why standard extensional probabilities encounter three fundamental problems: the problem of sample size, the problem of inclusion, and the problem of exceptions. These are discussed because their existence provides a rationale for the proposed model that will allow for their simultaneous resolution. I will suggest that pattern probabilities provide an improved criterion of adequate predication, and that quite plausibly, actual human probability

judgments may often approximate logical pattern probabilities as well. (Noteworthy here: if one is mainly interested in the technical realization of the model rather than in its rationale, one may wish to pass over the next sections and proceed directly to the model.)

(1) *The problem of sample size.* If ten events in X were observed, and A & B were observed all ten times (here “&” is used to describe single cell evidences, and “^” to refer to a conjunctive connective as a whole), the resultant belief in the hypothesis $A \wedge B$ would of course be stronger than had the observation been made only once. A direct application of extensional probability (ratio of confirmative to all cases in X), however, is not sensitive to different sample size; that is, $P_E(A \wedge B|X) = 1$ holds in both situations. This invites the use of a more refined measure, reflecting probabilities of probabilities, as suggested above, and shown in the proposed model.

(2) *The problem of inclusion.* Using extensional probabilities and a high probability criterion, a justified predication of, for example, “ravens are black and they can fly” entailed that one is at least equally justified to state that “ravens are black OR they can fly or both” ($P_E(B \wedge F|X) \leq P_E(B \vee F|X)$). This obtains, although the latter predication appears uninformative, if not misleading (cf. Grice’s principle of co-operation, 1975). Indeed, it is contradictory to the plausible goal of situation-characterization to use an uninformative, more general predication (such as a disjunction or even a tautology). Paradoxically, for any empirical situation in which one can use an AND-predicate, an extensional high probability criterion equally entitles us to use any predicate that is more general.²

An extensional account of probability judgments requires that the probability of a subset never exceeds that of its superset. Correspondingly, the *conjunction rule* states that a conjunction “ $A \wedge B$ ” is never more probable than either conjunct (“ A ” or “ B ”):

$$P_E(A) > P_E(A \wedge B), \quad P_E(A \wedge B) < P_E(B). \quad (1)$$

This rule has been called the most basic and universal law of probability theory (Kahneman & Frederick, 2002; Tversky & Kahneman, 1983). Standard inductive logic has used extensional probabilities, involving this rule (Carnap, 1952; Hempel, 1945; Reichenbach, 1935; Skyrms, 1986; see also Costello, 2005; Fitelson, 2006). Even non-standard accounts of probability, such as Cohen’s Baconian probabilities, Dempster–Shafer belief-functions, and types of multivalued or fuzzy logic did not deviate from this rule (Tversky & Kahneman, 1982, p. 90; Hájek, 2001, 2002).

Furthermore, the use of Bayes’ theorem alone does not provide any cure; for if we continue to use extensional probability, $P_E(A \wedge B|D) \leq P_E(A|D)$ remains valid (Fisk, 1996). Nor is it sufficient to interpret probability subjectively as belief, and then to update the probability of each cell individually; for adding the probability of one cell to another can only increase – never decrease – the probability of a union of cells: that is, $P_E(A \wedge B) + P_E(A \wedge B)$ merely yields $P_E(A)$.

The use of subjective *pattern* probabilities (P_P), in a broad sense, at least takes us one step further. Pattern $A \wedge B$ and pattern A may be seen as alternative (not nested) descriptions of an overall situation. When faced with a single A & B observation (data D), the likelihood of the connective $A \wedge B$, understood as an

explanatory propensity of the given data, is then $P(D|A \wedge B) = 1$. In contrast, if the data were produced by A or by $A \vee B$, lower likelihoods arise, since these hypotheses permit other outcomes as well. Assuming equi-probability of all possibilities allowed by a connective (cf. Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999), one obtains $P(D|A) = 0.5$ and $P(D|A \vee B) = 0.33$. Using Bayes’ theorem and a uniform prior, the posteriors of these pattern probabilities in the above example violate the conjunction rule: $P_P(A \wedge B|D) > P_P(A|D) > P_P(A \vee B|D)$. Note that we here deal with probabilities of patterns probabilities, given the data. Once again, in contrast, extensional probability (P_E) would be “1” for all three connectives. In effect, then, applying the conjunction rule on the level of probabilities of alternative hypotheses provides a “Bayesian Occam’s Razor”. In another debate, this has been advocated as “size principle” or “strong sampling” (Tenenbaum & Griffiths, 2001). It is an important step in solving the inclusion problem, and it arises naturally if one applies subjective Bayesian methods intensionally (rather than extensionally), taking the category size into account. Disappointingly, however, the idea itself does not solve the inclusion problem in the common case when the logical predications are noisy. Even if there are only a few exceptions this means that the rule is falsified and the more general hypothesis must be adopted. This may be a reason for the broadly accepted contention that neither frequentist nor subjectivist accounts of probability can provide a rational account of violations of the conjunction rule in human probability judgments (e.g., Fisk, 1996; Gigerenzer, 1998, 2000; Neace et al., 2008).

(3) *The problem of exceptions.* It is posited here that the inclusion-problem can only be solved together with the exception-problem. Correspondingly, the term “pattern probabilities” will be reserved as it relates to noisy-logical rules. Whereas predicate logic (employing only universal and existential quantifiers) suggests that general predications implicitly involve universal quantifiers (i.e., “all X are A and B ”), most material common-sense predications actually allow for exceptions (Schurz, 2001, see also Schurz, 2005). The statement “ravens are black and they can fly”, for example, cannot preclude the existence of ravens that are non-black (e.g., albino ravens, *Corvus alba*) or ravens that cannot fly. According to extensional probabilities (and even to a pattern-probability approach without noise), one here would have to assign the highest probability to a tautology such as “ravens are black or not, and they can fly or not”. Therefore the predication of a tautology based on the high probability criterion would be most valid. The proposed model of probability judgments of noisy-logical predications allows for a more reasonable probabilistic truth criterion.

4. A Model of the Bayesian logic of frequency-based inclusion fallacies

Building on the idea that “probability” is polysemous (Hertwig & Gigerenzer, 1999; Sloman & Over, 2003; see also Piaget & Garcia, 1991), a mathematical model for pattern-based probabilities of noisy-logical predications is proposed that allows for apparent violations of the conjunction rule. The proposed model may properly be termed *inductive Bayesian pattern logic of noisy-logical relations* (shortened to “Bayesian logic” or “Bayes logic” (BL) (von Sydow, 2009, 2007a, 2008)). The model allows for an improved high probability criterion for judgmental probabilities about noisy-logical hypotheses where the concern is with whole situations. It is a generative model that explains the occurrence of instances in a 2×2 table, based on underlying, logically described but probabilistic propensities, dispositions, or capacities (philosophically expressed: *potentia* vs. *actus*).

The model provides an alternative formalization to extensional probability—not replacing it, but supplementing it. On the level

² This problem may lie at the core of the paradoxes of implication in the narrower sense as well (e.g., Evans & Over, 2004). From “ravens are black and they can fly” it follows logically that the material implication “if ravens are black then they can fly” holds as well ($(B \wedge F) \leq (B \rightarrow F)$) (von Sydow, 2009). Since, however, the predication of conditionals raises more specific problems due to their possible counterfactual interpretation (Oberauer et al., 2007; Over, Hadjichristidis, Evans, Handley, & Sloman, 2007; Stalnaker, 1968), their possible incomplete representations (Johnson-Laird & Byrne, 2002), or their causal meanings (Hagmayer et al., 2007; Oberauer et al., 2007; Pearl, 2000; Sloman & Lagnado, 2005), this is not discussed here. One may alternatively explain the paradoxes of implication by specific properties of conditionals, but it is incontrovertible that, for example, the conjunction $\forall xB(x) \wedge F(x)$ is included within the disjunction $\forall xB(x) \vee F(x)$, although the latter may appear useless in the raven example.

of a description normally taken to be relevant (i.e., the relative frequency of actually observed cases falling under a rule), BL subscribes neither to extensionality nor to the conjunction rule (Eq. (1)). The corresponding non-additivity on the level of observed cases, however, results as an emergent property for BL. Basically, BL is likewise formalized in terms of standard Kolmogorov-probabilities, embracing additivity only at the different level of several noisy-logical hypotheses. In the next section sketches a rational analysis of the model.

4.1. Probabilities of hypotheses of noisy-logical predications: a rational analysis

According to the methodology outlined by Anderson (1990); see also Chater and Oaksford (2000), a rational analysis specifies an optimal solution to achieve the goals of a cognitive system within a particular environment (as opposed to specifying details of underlying cognitive processes). BL is formulated on this functional or *computational* level and does not address the cognitive implementation, or the *algorithmic* level (Marr, 1982).

The model is applicable in *environments* in which logical relations between properties normally do not hold deterministically (i.e., objective uncertainty) or in which observations may have been distorted (i.e., subjective uncertainty). Either of these criteria may be fulfilled in contingent predication, since empirical relations almost always allow for exceptions (e.g., the raven-example). In such environments, the introduced noise-factor is indispensable. Nonetheless, it is assumed that one has subjectively available data about the co-occurrence of binary properties represented in a 2×2 table.

Moreover, the model applies if the communicative situation in which a probability judgment is uttered recalls the *goal* of holistic rather than partial assessment (i.e., assessing a whole situation rather than piecemeal subsets). It evaluates which of several logical hypotheses is most likely to have produced the given data. As suggested in the Introduction, an important communicative rationale for pattern probabilities of noisy-logical rules is the need for a probability measure permitting maintenance of a high-probability criterion of predication by solving the problems of inclusion and exception. Additionally, the suggested model should yield probabilities for logical patterns, allowing for characterization by a single judgment of high probability. In contrast extensional probabilities require for characterization the specification of all four probabilities of a given co-variation matrix. Ultimately, it is plausible for some combination of evolution, cultural development, or individual learning to have produced a more efficient means for communicating about noisy-logical relations.

4.2. Main predictions

Before presenting technical details of the model I offer a number of “main” predictions derived from it. It should become clear that pattern probabilities of logical hypotheses — $P_P(A \circ B)$ — differ considerably from a direct application of extensional probabilities — $P_E(A \circ B)$. It is clear from the literature on probability-judgments that people sometimes ascribe to the word “probability” this standard statistical meaning, the relative frequency of a particular subset alone. Nevertheless, in situations addressing the probabilities of noisy-logical hypotheses and aimed at characterizing whole situations, BL should be applicable and may in fact predict systematic frequency-based violations of the conjunction rule. When these preconditions are met, BL yields a number of qualitative and quantitative predictions.

(1) Qualitatively, BL predicts that it should be possible to obtain a substantial portion of conjunction “fallacies”, even if one uses

frequencies as well as other factors proposed in the debate on the conjunction-fallacy (CF) as preventive factors. This prediction is interesting, since frequency format has often been found to reduce the rate of committed CFs (Fiedler, 1988; Hertwig & Gigerenzer, 1999; Reeves & Lockhart, 1993; Tversky & Kahneman, 1983)—which then gave rise to the idea that natural frequency representation generally reduced fallacious reasoning (Gigerenzer, 1991, 1996, 2000; Hertwig & Chase, 1998; Kahneman & Frederick, 2002, 2005; Mellers, Hertwig, & Kahneman, 2001; Messer & Griggs, 1993; but see also Sloman et al., 2003; Wedell & Moro, 2008). Other studies have shown similar “further factors” that reduce CFs. These include highly transparent within-subjects tasks (Gigerenzer, 1996; Kahneman & Frederick, 2002; Kahneman & Tversky, 1996; Lagnado & Shanks, 2002; Mellers et al., 2001); a rating response format (Hertwig & Chase, 1998; Hertwig & Gigerenzer, 1999, Study 4; Sloman et al., 2003; Wedell & Moro, 2008); clear set-inclusions (Agnoli & Krantz, 1989; Evans, Handley, Perham, Over, & Thompson, 2000; Moutier & Houdé, 2003; Neace et al., 2008; Over, 2004; Sloman & Over, 2003; Sloman et al., 2003; see also Evans et al., 2000; Johnson-Laird et al., 1999); absence of narratives (Kahneman & Frederick, 2002; Stolarz-Fantino, Fantino, & Kulik, 1996; Tversky & Kahneman, 1983; in contrast, see Gavanski & Roskos-Ewoldsen, 1991; Nilsson, 2008); and clarified logical formulations (Agnoli & Krantz, 1989; Hilton, 1995; Mellers et al., 2001; Messer & Griggs, 1993; Morier & Borgida, 1984; in contrast, see Sides et al., 2002; Tentori, Bonini, & Osherson, 2004). Such research has convincingly shown that some or all of the cues mentioned often reduce the portion of CFs. Nonetheless, I maintain this is partly due to eliciting an extensional interpretation of probability. Moreover, BL predicts that these cues are in fact not sufficient to elicit extensional reasoning. Finally, although this paper does not focus on the evaluation of specific cues, it does demonstrate that, when the goal is to make probability-judgments about sentences characterizing an overall situation, CFs can be elicited even if all of the above cues are applied simultaneously.

(2) Quantitatively, BL makes a number of predictions for the probability judgments modeled, based on the observed frequencies in a contingency table.

(a) BL predicts not only CFs but double CFs as well as the quantitative conditions of their occurrence. Study 1 will explore which conditions lead participants to expect conjunctions to be more probable than *both* conjuncts together with effects of negations.

(b) BL predicts that sample size variation may have a differential effect on estimated probabilities, even if extensional probabilities are held constant. Studies 2a and 2b provide a first test of such differential sample size effects.

(c) BL predicts internal and external pattern-sensitivity effects. For instance, according to pattern-sensitivity, probability judgments should vary in a specific way if the distributions of confirmatory (or disconfirmatory) cases is varied inside (or outside) a given set, even if the extensional probabilities remain constant. These effects provide a good example of the mentioned communicative goal linked to BL. For instance, in considering a previously unknown species *X*, one may focus on two properties — warm fur (*A*) and good meat (*B*) — and their logical relationship; whereupon one may wish to communicate one’s observations. Sample scenarios may be formulated thus: 18 *A* & *B*, 5 *A* & *non-B*, 6 *non-A* & *B*, and 7 *non-A* & *non-B* [18, 5, 6, 7]; or, with a different data-vector: 18, 15, 1, 2. In which scenario would the hypothesis “animals of species *X* provide good meat AND warm fur” appear more probable? According to an intuitive use of BL, the naive probability of the AND-hypothesis should be higher in the former scenario than in the latter. According to extensional probability, however, the relative frequency leads to identical judgments: $P_E(A \wedge B) = 0.5$.

Moreover, extensional probability would require four conjunctive probability judgments – one for each logically possible subset – to communicate the pattern of probabilities. Logical pattern probabilities allow for an overall characterization by assigning a high probability to a single sentence. Such effects of internal and external pattern sensitivity are explored in Study 3.

Additionally, some further predictions (*d, e, f*, not investigated here) should be mentioned:

(d) BL is not limited to summary data but can be applied as well to single data-points or trial-by-trial learning of logical relations. Since BL uses occurrences in contingency tables as an input, single observations can be modeled parsimoniously as a contingency table with a set-size of 1. Trial-by-trial learning could simply use the posterior probabilities as a new prior probability distribution for each trial. This can only be done if no specific effects of attention or memory decay are assumed (for interesting additional effects of subjective model construction, cf. Betsch & Fiedler, 1999). If any such effects play an important role, it would be necessary to assess the subjectively represented frequencies in order to apply the model properly.

(e) BL postulates an impact of prior probabilities. For the experiments presented here, flat prior distributions will be assumed, due to the laboratory settings used. In other contexts, BL predicts the effects of priors, particularly for very low sample sizes (Kahneman & Frederick, 2002; Kahneman & Tversky, 1973). More specifically, the prior probabilities of particular noise levels may affect the posterior probability of different logical hypotheses.

(f) Finally, BL predicts the systematic occurrence of other logical inclusion “fallacies” in frequency-based tasks, including the “disjunction fallacy” (Reeves & Lockhart, 1993). Thus BL goes far beyond the previously studied phenomena in specifying a system of rational logical inclusion fallacies (for first corroborative results see von Sydow, 2009).

In sum, BL makes a number of qualitative and quantitative predictions concerning the occurrence of frequency-based conjunction fallacies (and other inclusion fallacies), where the concern is probability judgment and the intuitive logical evaluation of an overall situation.

As mentioned above, a standard (extensional) Bayesian approach cannot rationally explain violations of the conjunction rule (Bar-Hillel, 1991; Fisk, 1996; Gigerenzer, 1991). Yet, consonant with a renaissance of Bayesian models in psychology (Chater & Oaksford, 2008; Chater, Tenenbaum, & Yuille, 2006; Hahn & Oaksford, 2007; Kruschke, 2008; Oaksford & Chater, 2007, 1994; Tenenbaum & Griffiths, 2001; Tenenbaum, Griffiths, & Kemp, 2006; see also Nelson, 2005), the development of the idea of logical pattern probabilities incorporating the use of different noise levels permits the reconstruction of deviations from the conjunction rule as rational Bayesian answers.

4.3. The model

BL formulates posterior probabilities of noisy-logical explanations (themselves each involving patterns of probabilities), given observed frequency data in a contingency table. It is assumed that the data could be understood to be generated by propensities described in a logical but probabilistic way. Each noisy-logical hypothesis is taken to have generated the data and is a potential explanans of an observed situation. These hypotheses are represented by probability tables (PTs; here, 2×2 -tables), each of which corresponds to an ideal logical truth table plus a degree of noise. The model first generates all possible explanatory logical rules; then it produces possible noise levels for each rule. Based on these noisy-logical representations, one can determine the likelihood that the data have been produced by each possible underlying logic-noise pattern (a hypothesis). Using a flat prior

distribution for these hypotheses, one can employ Bayes's theorem to calculate the posterior probabilities for each logic-noise combination. Then, to determine the overall posterior probability of a single logical hypothesis, the posteriors of a particular logical hypothesis are summed up over all modeled noise-levels.

On the logic side, BL is concerned with all sixteen possible dyadic connectives between two atomic propositions *A* and *B* (Frege, 1879)—AND, OR, EITHER OR, and so forth: $A \circ_l B$ (with the index *l* running from 1 to 16). Nevertheless, the focus here will be on the three connectives “*A* and *B*”, “*A* (which are *B* or not-*B*)”, and “*B* (which are *A* or not-*A*)”, since they were particularly relevant in the conjunction fallacy debate considered in the General Discussion. In what follows, BL will be confined to dyadic logic of probabilistic logical relationships between *A* and *B* in a universe of discourse *X*, with analogous noisy pattern-logics being constructable for relationships of higher arity.

To begin, the model should be developed in steps, as shown below.

(1) *Inputs*. The main inputs for Bayesian logic are frequencies as represented in a 2×2 -contingency table resulting from *N* independent observations: $x_a + x_b + x_c + x_d = N$, with $N \geq 0$ (see Table 2). The contingency-table presupposes a given universe of discourse *X* (e.g., “ravens” or “graduates of the Linda-school”). To apply the model without modifications, the sampling-process producing the data should be unbiased (von Sydow, 2008; cf. Fiedler, 2000), whether the data is directly observed or retrieved from memory. Since such evidence combines information about effect and sample-size, we are concerned with “natural frequencies”, defined to report the final “tally of a natural sampling-process” (Gigerenzer, 1998, p. 13; Gigerenzer & Hoffrage, 1995; Gigerenzer & Hoffrage, 2007).

As a second input, BL assumes prior probabilities for the tested logical hypotheses at each noise level modeled. In the studies, a uniform, uninformative prior distribution was employed for all combinations of hypotheses and noise levels (cf. von Sydow, 2007a, 2008, 2007b, for possible extensions).

(2) *Probability tables*. Similar to other kinds of multivalued or fuzzy logics, BL replaces the two values, *true* (*T* or 1) and *false* (*F* or 0), of bivalent propositional logic, by values within the interval $[0, 1]$ (as used for probabilities). BL formulates specific probability tables (PTs) as well as probabilities on a second level for each of these probability tables. PTs are probabilistic analogues to logical truth tables, which are treated as generative structures that may have created the data. Since the concern is with dyadic relations, each PT here is based on tuples of four probabilities: $P_E(A \& B) + P_E(A \& \neg B) + P_E(\neg A \& \neg B) + P_E(\neg A \& B) = p_a + p_b + p_c + p_d = 1$.³ It should be noted that these PTs do not directly summarize the observations; rather they are ‘a priori’ explanatory constructs of the model, needed to instigate this kind of inductive logic. Each PT as a whole, however, has a probability that changes when tested against data, yielding a new posterior.

PTs are probabilistic descriptions of hypothetical logical propensities that can give rise to certain empirical attribute combinations. The PTs used make up a subset of all possible PTs selected to model pattern probabilities of noisy-logical rules. In constructing these PTs, two main assumptions are made: first, idealization; and second, uncertainty.

(3) *Assumption of idealization*. Logical truth tables concern only the truth or falsity of cases—what is allowed and what is forbidden. They do not provide probabilities for the allowed cases. In fact, the probability of a confirmatory subclass may be zero. Nevertheless, since PTs are taken here to represent

³ This equation does not hold for PTs based on the logical falsum, whose probabilities are determined by noise only.

generative structures, it would be absurd to use a probability of zero for a subclass, since this would rule out completely that this option generated the data. According to the idealization-assumption, a hypothetical ideal logical connective is assumed to have an equal probability distribution for all true cases (Johnson-Laird et al., 1999; see also, in a context not dealing with logics or exceptions, Tenenbaum & Griffiths, 2001). The connective corresponds to the question of which (ideal) logical explanation could have created a given data-set. Moreover, the assumption follows naturally from the principle of indifference, as long as no additional knowledge about sampling constraints or distorting factors is available. In the case of deterministic logical relationships (no noise, $r = 0$), the probability of a cell of a PT being logically “false” remains zero, whereas the ideal probability of a cell corresponding to any “true” case of a connective (in the absence of noise) is 1 divided by the number ν of true cells T given by this connective l , or: $P_{PT}(T|r = 0) = 1/\nu(T_i) = t_i$. For the conjunction “A AND B”, we obtain $t_1 = 1$; for the affirmation “A”, $t_1 = 0.5$; and for the (inclusive) disjunction “A OR B”, $t_1 = 0.33$. This step turns deterministic truth tables into (deterministic) probability tables. The measure t_i is not a free parameter; rather, it follows from the idealization-assumption. It must be noted, however, that the observed data generated by a logical PT can be more irregular.

(4) *Assumption of uncertainty.* In an uncertain world, a formalization of noise (error, nuisance, uncertainty, or randomness) is needed in order to develop a full idea of probabilistic logical PTs. A deterministic PT (with $r = 0$), as in the case of a truth table, could be discarded by a single disconfirmatory observation (a falsification; cf. Popper, 1934/2005). For each connective l , therefore, several PTs are constructed, each with a different noise level r_j . (Note that only eleven noise levels were modeled here; thus r_j was varied by increments of 0.1.) The assumption of uncertainty formalizes a general overall level of uncertainty r (with $0 \leq r \leq 1$) for all cells of a given logical PT, assuming that we have no knowledge about specific noise factors.⁴ A resulting connective-noise combination (a PT) is an ideal logical but noisy explanans and may have generated the data. Each PT has an additional conjoined probability: $P(PT_{l,j}) = P_p(A \circ_l B, r_j)$. As prior probability, the probabilities of a connective-noise combination or of overall noise levels may be fixed by prior knowledge. Here flat priors are assumed, however, and the model is used to calculate the posterior probability of each such combination ($P_p(A \circ_l B, r_j|D)$) from the data itself.

Formally, r is the portion of the PT’s generative probability distributed equally over all cells of a connective. For PTs with no noise ($r_j = 0$), the probability of a logically false case producing an observed case is $P_{PT}(F) = 0$. Hence, after only a single observation corresponding to a forbidden cell, does the posterior of the PT become zero (a falsification). In contrast, for PTs with maximal noise ($r = 1$), the cell-probabilities converge at the *convergence-probability*, here generally set to 0.25. No additional parameters are introduced that could be suitable in more complex situations (von Sydow, 2007a, 2008, 2007b).⁵ Hence the cell probabilities of each PT can be represented as a mixture of (1) a uniform distribution over the intension of the truth table of a connective, and (2) a uniform noise distribution over all four cells.⁶ As an example, when

the connective of the rule is “OR” (inclusive disjunction), the 2×2 probability table reads:

$$P_{T_{OR,j}} = (1 - r_j) * \begin{bmatrix} 1/3 & 1/3 \\ 1/3 & 0 \end{bmatrix} + r_j \begin{bmatrix} 1/4 & 1/4 \\ 1/4 & 1/4 \end{bmatrix}. \tag{2}$$

And when the connective of the rule is “AND”:

$$P_{T_{AND,j}} = (1 - r_j) * \begin{bmatrix} 1/1 & 0 \\ 0 & 0 \end{bmatrix} + r_j \begin{bmatrix} 1/4 & 1/4 \\ 1/4 & 1/4 \end{bmatrix}. \tag{3}$$

For each connective, different noise levels (r_j) can be represented. Although the noise levels approximate a perfect (continuous) representation only when their number approaches infinity, for our psychological purposes here only eleven equidistant discrete levels of r were used.

(5) *Likelihoods.* The formalization of the novel pattern-interpretation of (noisy) logical connectives results in a two-dimensional field of PTs (connectives \times noise), each with its conjoint probability. The subsequent steps combine standard procedures of Bayesian statistics with these non-standard representations of distributions of noisy-logical relationships between two predicates.

The formalization of the PTs allows a calculation of the likelihoods of observed data patterns given a connective-noise hypothesis $P(D|PT_{l,j}) = P(D|A \circ_l B, r_j) = P(x_a, x_b, x_c, x_d|p_a, p_b, p_c, p_d)$. The multinomial distribution provides the discrete probability-distribution for obtaining a particular pattern of data (D) in a sample of independent observations (when $\Sigma x = N$), given a hypothesis (a PT) with the respective probabilities p_a, p_b, p_c , and p_d (with $0 \leq p_m \leq 1$). For a given data pattern and each $PT_{l,j}$ (that is, each PT specified by the model), $P(D|PT_{l,j})$ is calculated thus:

$$P(D|A \circ_l B, r_j) = \binom{N}{x_a x_b x_c x_d} p_a^{x_a} p_b^{x_b} p_c^{x_c} p_d^{x_d}. \tag{4}$$

(6) *Bayes’s Theorem.* This theorem is used to calculate the posterior probabilities ($P(PT|D)$) of each connective-noise hypothesis, given the observed pattern of data.

$$P_p(A \circ_l B, r_j|D) = \frac{P(D|A \circ_l B, r_j)P(A \circ_l B, r_j)}{P(D)}. \tag{5}$$

The normalizing probability ($P(D)$) is obtained by summing up the weighted likelihood of the data under all connective-noise hypotheses.

$$P(D) = \sum_l \sum_j P(D|A \circ_l B, r_j)P(A \circ_l B, r_j). \tag{6}$$

(7) *Integration rule.* The logical pattern probability of a dyadic hypothesis, $P_p(A \circ_l B)$, is calculated by summing up the posterior probabilities of all corresponding $PT_{l,j}$ over all noise levels j ⁷:

$$P_p(A \circ_l B|D) = \sum_j P_p(A \circ_l B, r_j|D). \tag{7}$$

The resulting probability mass function provides us with the desired posterior probabilities for each noisy-logical hypothesis (noisy-logical pattern probabilities).

(8) *Comparative probability judgments.* Additionally, one may well be interested in ratios of the posterior probabilities rather than in their absolute differences. This becomes significant

⁴ An alternative formalization may be reasonable as well, leading to similar predictions here (cf. von Sydow, 2009). Moreover, noise might be understood as specific causes of noise, which may be captured within the current model by summing up the probabilities of all alternative hypotheses.

⁵ Generally, the assumed probability of a false cell in a $PT_{l,j}$ is formalized by $P_{PT}(F) = r * c$ (with $0 \leq P_{PT}(F) \leq c$); and that of a true case, T , by: $P_{PT}(T) = (1 - r)t_1 + r_c$, with $c \leq P_{PT}(T) \leq t$.

⁶ I owe this didactically improved formulation to a comment by Professor John Kruschke. It is equivalent to the one proposed in von Sydow (2007a,b): $P_{PT}(T) = (1 - r)t_1 + r_c = t_1 - r(t_1 - c)$.

⁷ There may be other weighting functions and integration rules as well. Pragmatically, one may, for instance, favor low over high noise level hypotheses and weight r values by their reciprocal, $1/r$, normalizing this by the overall sum. It should be noted, however, that this would not be a purely epistemic model, concerned with truth alone, but one concerned with epistemic utility.

when probability-differences close to zero nevertheless refer to large ratios. Thus it may be more appropriate to model human judgments using either a logit-transformation of the probability scale as output (von Sydow, 2008), or a directly comparative measure for comparing each hypothesis (e.g., log posterior odds; see Anderson, 1990; Griffiths & Tenenbaum, 2005): $\ln(P_P(A_{O_1}B|D)/P_P(A_{O_2}B|D))$. If several comparisons are involved, for each hypothesis one may use the average of all log posterior odds. In the studies discussed here, where such values are shown, logit-values and log-posterior odds yield very similar predications.

In sum, this formalization of inductive Bayesian logic yields a rational model for pattern probabilities ($P_P(A_{O_1}B|D)$) that predicts the systematic occurrence of CFs and double CFs under specified qualitative and quantitative conditions. The model as presented here has no free parameters (given the use of flat priors and ignoring the model variants in Step 8). Based on this formalization, even if BL were only applicable on an ordinal level (see Stewart, Brown, & Chater, 2005, Study 3), it is possible to specify the predictions for our studies in a precise way.

5. Related models

In this section some models are briefly discussed that are similar in spirit but different in subject matter as well as in execution. Models explicitly addressing the conjunction fallacy debate will be considered in the General Discussion.

BL differs mathematically from models using the term “Bayesian logic” in a different or more general way (Busemeyer, Wang, & Townsend, 2006; Carbonetto, Kisyński, de Freitas, & Poole, 2005; Milch et al., 2006; Romeyn, 2005; von Sydow, 2006). Hence, in contexts in which a confusion of terms is unlikely, one can refer to the proposed model briefly as “Bayesian logic”.

Earlier studies have mentioned that neither non-standard accounts of probability theory (such as Baconian probability, belief-functions or fussy logic, see Tversky & Kahneman, 1982, p. 90; Hájek, 2001, 2002) nor classical accounts of inductive logic can explain violations of the conjunction rule (Carnap, 1952; Hempel, 1945; Reichenbach, 1935; Skyrms, 1986; see also Costello, 2005; Fitelson, 2006). BL also differs from broadly related Bayesian accounts in the Raven paradox debate (e.g., Nickerson, 1996; von Sydow, 2006; Vranas, 2004), and from models in AI and machine learning at the intersection of logic and probability (e.g., Carbonetto et al., 2005; De Raedt & Kersting, 2003; Milch et al., 2006). These approaches addressed different problems and did not constitute a pattern-sensitive inductive Bayesian logic integrating over noise levels.

As mentioned, the model is in line with the size principle advocated by Tenenbaum and Griffiths (2001) (for a logical albeit non-Bayesian context, see Johnson-Laird et al., 1999), but the idea was not developed in the context of logical propositions, nor was the problem of exceptions for logical predications resolved (violations of rules were treated as falsifications). Notably, however, two recent models at the intersection of logic and probability-theory have explicitly dealt with probabilities of logical relations as well as exceptions: Oaksford and Chater (2003, information-gain model) and Goodman, Tenenbaum, Feldman, and Griffiths (2008, categorization-model).

In particular, Oaksford and Chater’s model of optimal information gain (OIG) in Wason’s four-card selection task (cf. Tentori, 2002; Oaksford & Chater, 2007, 1994, 2003; von Sydow, 2006, 2004) seems to anticipate aspects of BL in characterizing a connective by probability patterns and in allowing for exceptions (Oaksford & Chater, 2003). Moreover, the model directly influenced work leading to the BL model (von Sydow, 2006). Nonetheless, BL does differ from OIG in essential respects. First, the OIG model has *prima facie* a different field of application, being concerned with

optimal data selection. Moreover, it deals with material implications alone. Although I have suggested that this model may be extended to other connectives, such as disjunctions, there is no direct way to extend it, for instance, to conjunctions (von Sydow, 2006). Second, the OIG model assumes given information about the marginal probabilities of two predicates but *no* knowledge of their co-occurrence. In contrast, BL assesses fully available frequency information. Furthermore, OIG specifies only one noise level, not all possible noise levels. It is crucial that, although employing one type of logical PT, OIG deals with a basically extensional understanding of connectives. Probability tables are constructed only as a side effect resulting from assumed constraints, not with reference to ideal logical patterns. Although empirical tests of the OIG model in Wason selection tasks revealed mixed results, some studies found supporting evidence only when these constraints were actually in place (von Sydow, 2006, 2004). In contrast, Bayesian logic concerns logical pattern probabilities, presuming knowledge about the *complete* contingency table and *no* sampling constraints (cf. Fiedler, 2000).

Goodman et al. (2008) recently proposed a rule-based categorization model using a logical grammar in disjunctive normal form (DNF). Since both this model and BL (von Sydow, 2007a, 2008, 2007b) use logical rules and exceptions, they may appear similar. Yet there are essential differences: on the one hand, DNF has as its goal to find, via features or feature-combinations, the best definition for two alternative categories (categorization-task). That is, feature-combinations that occur in two categories (ambiguous feature combinations) would not be relevant. On the other hand, BL is concerned with the predication of the logical relation of some possible features in question, given a particular category. In addition to this difference, the likelihoods in the DNF model decrease extensionally with the number of assumed outliers. DNF favors simpler models (even if less adequate) by the use of a syntactic complexity bias (using a disjunctive normal form) that is reduced in successive learning. This is a reasonable formalization for this task because the model is formulated to discriminate increasingly between different alternative categories in a perfect way (without outliers); nonetheless, it differs from the task modeled by BL and from the prediction of no erosion of assigning pattern probabilities involving CFs with increased sample size.

In sum, the outlined models clearly display differing goals and technique. Despite this fact, however, BL shares with these models the overall aim to bridge the often claimed abyss between logical and probabilistic representations, or between Boole and Bayes.

6. Testing Bayesian logic

In three studies, central qualitative and quantitative predictions of BL were investigated, which in their combination (as will be shown later) cannot be explained by any other theory of the CF.

Qualitatively, BL predicts the possibility to elicit CFs even under conditions using many extensional cues simultaneously, if the concern is alternative hypotheses about whole situations. In all experiments, conditions were used under which most other theories would *not* predict CFs. If this led to CFs, the result would go beyond even the most startling recent demonstrations of CFs (e.g., Lagnado & Shanks, 2002; Sloman et al., 2003; Tentori et al., 2004). (Note: Traditional accounts of CFs will be addressed afterward in the General Discussion.)

In particular, the goal of the studies was to test three previously mentioned central *quantitative* predictions: the role of negations and the conditions of double CFs (Experiment 1); differential sampling size effects (Experiments 2a and 2b); and internal and external pattern sensitivity (Experiment 3). The experiments concentrated on the predictions of BL, testing them against an extensional understanding of probability.

6.1. Experiment 1: double conjunction fallacies and negation

Experiment 1 had three purposes—in ascending order of interest: (1) to explore qualitatively whether probability judgments deviate from the norm of extensional probability even with many extensional cues in place; (2) to control for possible modulating effects of negations and presentation position; and finally, but most important, (3) to elucidate the quantitative conditions under which double CFs occur.

We used a scenario reminiscent of the classical CF-task, the *Linda-task* (Tversky & Kahneman, 1983). Instead of a story about one person (“Linda”), with probability judgments about an individual’s attributes, the tasks concerned graduates of schools (e.g., the “Linda-school”), with probability judgments about group attributes. According to the conjunction rule a probability of the affirmation “Linda is a bank teller” ($P(B)$) can never be less probable than the conjunction “Linda is an active feminist and a bank teller” ($P(A \wedge B)$). Participants often violated this rule in classical Linda-tasks when ranking the probabilities of the conjunction higher than the affirmation after Linda had been described as feminist. Here we use explicit frequency data instead of single-event narratives. Additionally, the task uses no distractor properties, clear set inclusions, within-subjects comparisons, and formulations of the hypotheses that exclude misunderstanding. Despite these extensional cues, in a context that suggests an interest in an overall evaluation of a situation BL predicts a relevant portion of non-extensional answers.

The ways the data patterns were linked to bank teller or (active) feminist were counterbalanced by using four rotations of the input matrix of each pattern type (see Fig. 1(A)). Moreover, negated hypotheses were investigated. Taken together, this allowed controlling for possible effects of affirmations versus negations in a setting with transparent frequencies (cf. Hattori & Oaksford, 2007). Moreover, it allowed control over possible misinterpretations of conjunctions as disjunctions or implications (Hertwig et al., 2008; Mellers et al., 2001). Additionally, we varied the presentation order, since the theories of inverse probability and support would predict that this factor is relevant for CFs (see the General Discussion).

The main goal of Experiment 1 was to investigate (a) the occurrence of frequency-based *double CFs*; and (b) whether they occur under the *quantitative conditions* predicted by BL. Double CFs are empirically measured probability judgments (P_{emp}) involving $P_{\text{emp}}(A) < P_{\text{emp}}(A \wedge B)$ and $P_{\text{emp}}(A \wedge B) > P_{\text{emp}}(B)$; or $P_{\text{emp}}(A) \leq P_{\text{emp}}(A \wedge B)$ and $P_{\text{emp}}(A \wedge B) \geq P_{\text{emp}}(B)$. The former are *strict* double CFs. Research on representativeness has so far concentrated on single CFs (such as $P(\text{bank teller}) < P(\text{active feminist} \wedge \text{bank teller})$) and on situations triggering a prototype related to one of the components (e.g., active feminist) (Neace et al., 2008; Thüring & Jungermann, 1990; Tversky & Kahneman, 1983). Often the conjunctions have not been analyzed together with *both* conjuncts. Only a few approaches have explicitly predicted double CFs (e.g., Costello, 2005; Yates & Carlson, 1986); and in only a few traditional Linda tasks (without full frequency information) have small portions of double CFs been found (Costello, 2005; Hertwig & Chase, 1998; Reeves & Lockhart, 1993; Yates & Carlson, 1986; cf. Tentori et al., 2004). Even Lagnado and Shanks (2002), who varied learned frequencies of different logical subsets in more complex settings, did not assess double CFs. In difference, here double CFs are investigated while providing full frequency information, including the conjunction itself (see also von Sydow, 2007a, 2008, 2007b).

Experiment 1 explores the *quantitative* conditions under which frequency-based double CFs occur. The four types of frequency patterns explored all have a modal conjunctive frequency in the same logical cell (see Fig. 1(A)). One may either derive the

Table 3

Presentation orders and resultant presentation positions of schools in Experiment 1 (cf. Fig. 3).

Order	t1	t2	t3	t4	Order	t1	t2	t3	t4
Order 1	1a	2b	3c	4d	Order 9	3a	1b	4c	2d
Order 2	1c	2a	3d	4b	Order 10	3c	1a	4d	2b
Order 3	1b	2d	3a	4c	Order 11	3b	1d	4a	2c
Order 4	1d	2c	3b	4a	Order 12	3d	1c	4b	2a
Order 5	2a	4b	1c	3d	Order 13	4a	3b	2c	1d
Order 6	2c	4a	1d	3b	Order 14	4c	3a	2d	1b
Order 7	2b	4d	1a	3c	Order 15	4b	3d	2a	1c
Order 8	2d	4c	1b	3a	Order 16	4d	3c	2b	1a

predictions of BL directly from the specified pattern probabilities (Fig. 1(B); see BL, Step 7) or else assume that people do not represent differences but rather proportions of these probabilities (Fig. 1(C), averaged log-odds; see BL, Step 8). Note that all hypotheses have a non-zero posterior probability. It is an advantage of the latter measure, showing the average log-odds of pattern probabilities, that allows representing the relative differences between low-probability hypotheses even if one hypothesis is dominant (Fig. 1(B) and (C)). In either case, BL predicts dominant double-CF effects only for two of these patterns (Pattern Types 1 and 2). Pattern Type 1 investigates whether, as a minimum criterion, double CFs occur with highly frequent conjunctive cells (with an extensional probability of $P_E(B\&F) = 0.8$). Pattern Type 2 explores whether they also occur in cases with a conjunctive probability of less than 0.5 (where $P_E(B\&F) = 0.4$). Even here, BL predicts dominant double CFs, although BL allows finding reduced confidence in such judgments (see ordinates in Fig. 1(B)). For Pattern Types 3 and 4, there should be fewer double CFs, although in the case of Pattern Type 3 the extensional probability of the conjunction is now even higher than in Pattern Type 2; and in Pattern Type 4 it comes closer to an AND-pattern. Hence, any measure demanding only a high extensional probability of a conjunction for double CFs to occur would lead to different predictions. Likewise, these predictions are not made by traditional theories of the CFs (see General Discussion).

Method

Participants.

Sixty-four students of the Georg-August-Universität Göttingen volunteered to participate, in return for a small chocolate bar. Additionally, they could enter a lottery as recompense.

Materials and procedure.

A mixed repeated-measures design was employed, with each participant receiving a booklet (in German) giving a general introduction, followed by four judgment tasks and a final questionnaire. The frequencies for each task were derived from the sixteen frequency patterns used, resulting from four rotations of the four pattern types introduced (Fig. 1(A)). The presentation order combined two Latin Squares so that no participant got the same rotation or pattern type twice (Table 3). The Latin Squares also secured that the patterns and rotations occurred equally often in all four positions, and that the types succeeded each other only once. This resulted in sixteen presentation orders (Table 3). Participants were randomly assigned these orders, with each being shown to four participants.

The tasks were presented in succession, one per page. Participants were informed to take an interest in an overall evaluation of the school shown (e.g., the Linda school). They should place a check-mark next to the hypothesis that *intuitively* appeared most probable, based on the sample. They were informed that none of the hypotheses presented was 100% valid.

The frequency information was presented to the participants in a contingency table. The table’s marginals were labeled “bank teller”, “no bank teller”, “feminist”, “no feminist”; and the

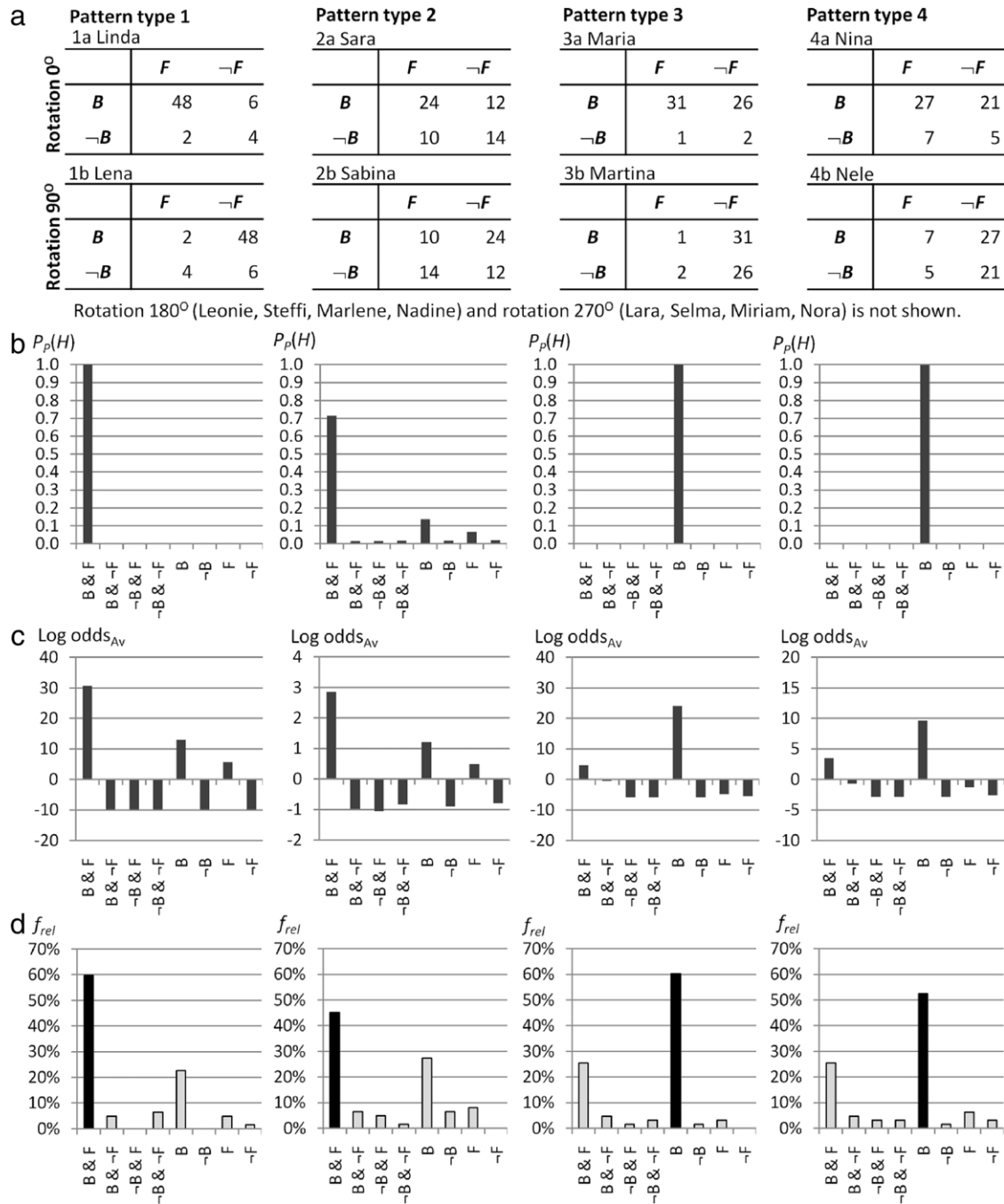


Fig. 1. Data patterns, predictions and results of Study 1. (A) The tables show the data patterns used, and their rotations, in the different schools. The other rotations are designed analogously. (B) The posteriors of the pattern probabilities of noisy-logical hypothesis, $P_p(H) = P_p(B \circ_i F|D)$ (BL, Step 7). (C) Averaged log-odds of the pattern probabilities (BL, Step 8). (D) Relative frequency of the selection of different logical hypotheses judged most probable for different types of patterns (recoded after conflating over Rotation and Position).

cells, “bank teller & feminist”, “bank teller & no feminist”, and so forth. This was done to clarify set inclusions and prevent misunderstanding.

Participants were provided with eight hypotheses regarding the Linda-school students, who after graduating generally became: (1) “bank tellers and at the same time feminists” ($B \& F$); (2) “bank tellers and [...] no feminists” ($B \& \text{Non-F}$); (3) “no bank tellers and [...] feminists” ($\text{Non-B} \& F$); (4) “no bank tellers and [...] no feminists” ($\text{Non-B} \& \text{Non-B}$); (5) “bank tellers” (B -hypothesis); (6) “no bank tellers” (Non-B); (7) “feminists” (F); and (8) “no feminists” (Non-B).

The eight hypotheses allow for investigating whether the use of negations leads to asymmetries. Notably, the used formulation excluded an interpretation of conjunctions as disjunctions (using the formulation “and at the same time”; Mellers et al., 2001; see also Hertwig et al., 2008), as well as the interpretation of “A” as an ellipsis for “A and non-B”, since the latter hypothesis is explicitly stated (Agnoli & Krantz, 1989; Messer & Griggs, 1993; Morier & Borgida, 1984; Tentori et al., 2004; Wedell & Moro, 2008). (For full instructions to the task, see Appendix A.)

Results.

Of the 256 data points, six were excluded from analysis for formal reasons (empty cells or multiple selections). Fig. 1(D) shows

the percentage of occasions a particular logical hypothesis was deemed most probable in the four pattern types, conflating over the factors Rotation and Presentation order/position. The re-coding reversed the rotations (e.g., actual *B* & *non-F*-, and *non-F*-choices in 90° rotations are re-coded as *B* & *F*- and *B*-choices).

In Pattern Type 1, 60% of the choices corresponded to the predicted conjunctions (e.g., *B* & *F* in 0° rotations), all involving double CFs. The second-highest group, 22%, selected a particular single affirmation or negation (e.g., *B* in 0° rotations, *non-F* in 90° rotations, and so forth), corresponding to the main extensional prediction—the second most probable hypothesis according to BL (see Fig. 1(C)). The remaining 18% were distributed across the other six hypotheses. For Pattern Type 2, the same general predictions held. The largest portion of re-coded choices, 45%, corresponded to the predicted conjunctions, with 27% to the extensionally predicted affirmations/negations (also the second most probable pattern probability), and 27% to the six other hypotheses. As predicted, the responses for Pattern Types 3 and 4 were the reverse of those for 1 and 2: 25% (in both 3 and 4) chose the focal AND hypotheses (corresponding to the second highest pattern probability); 60% (in 3) and 52% (in 4) chose the main affirmation/negation hypotheses (e.g., *B*), predicted here by both BL and extensional probability; and finally, 14% (in 3) and 22% (in 4) selected other hypotheses.

If participants misunderstood conjunctions as disjunctions, they extensionally should have answered, for instance, “*B* & *non-F*,” in 0° rotations of Pattern Types 1, 2, and 3. On the contrary, however, summing up the extensionally predicted selections under the hypothesis that an AND-formulation is misunderstood as a logical disjunction, these selections occurred significantly less frequently than by the chance-probability of one eighth (testing against all alternative selections with an a priori probability of seven eighths ($\chi^2(1, n = 187) = 8.75, p < 0.01$)). Notably, the main predictions of BL, given that grammatical conjunctions had been interpreted as logical conjunctions (the black columns in Fig. 1(D)), occurred significantly above chance level (always with $p < 0.00001$).

Comparing Pattern Types 1 and 2 (predicted conjunctions) with Types 3 and 4 (predicted affirmations/negations), the portion of (re-coded) AND-selections (vs. the portion of all non-AND-selections) clearly differed in the expected direction: $\chi^2(1, n = 250) = 19.22, p < 0.00001$. Even for the strictest possible single comparison – between Pattern Types 2 and 4 – this portion became significant: $\chi^2(1, n = 125) = 5.35, p < 0.05$. Fig. 1(D) seems to reveal a reduced portion of AND-selections (vs. non-AND-selections) between the two AND pattern types (Types 1 and 2). Such difference, which had not been predicted, need not be inconsistent with BL (Fig. 1(B) and (C)). In any case, the difference did not reach significance: $\chi^2(1, n = 124) = 2.62, p = 10.55$. Comparing the portions of AND-answers in Pattern Types 3 and 4 confirmed that there was in fact no difference.

Finally, selections of predicted affirmations/negations (re-coded, e.g., *B* in 0° rotations) versus all other selections differed, as expected, between Types 1 and 2 on the one hand and Types 3 and 4 on the other: $\chi^2(1, n = 250) = 25.43, p < 0.00001$. Furthermore, the most critical single comparison of this kind, between Types 2 and 4, was significant: $\chi^2(1, n = 125) = 8.49, p < 0.01$. The portion of *B*-selections (after re-coding) did not differ reliably between the two AND-pattern types ($\chi^2(1, n = 124) = 0.39, p = 0.53$), nor between the two *B*-pattern types ($\chi^2(1, n = 126) = 0.81, p = 0.36$).

Additionally, log-linear analyses, involving the factors “Pattern type × Rotation × Position × Possible answers”, suggest that participants’ selections of the most probable hypotheses were predominantly affected by Pattern type, and not by Rotation or Position (see Appendix B).

Discussion of Experiment 1.

(1) The results of Experiment 1 provide an existential proof of double CFs becoming dominant choices, under strict conditions, using natural frequency information, a contingency-table task, within-subjects comparisons of hypotheses, clear set inclusion, clarified hypotheses formulations, and no confounding variables. This shows that these cues, normally assumed to elicit correct (extensional) answers, need not be sufficient to elicit extensional reasoning.

(2) The experiment suggests that positive and negative evidence were treated symmetrically. The factor Rotations had no significant impact. This may have been due to the task’s deliberately low memory demands. That is, in more complex contexts, the representation of positive or negative cases could well differ (Hattori & Oaksford, 2007; Johnson-Laird et al., 1999), and then the predictions of BL, if based on subjective frequencies, would differ as well. All in all, the results corroborate that BL may *ceteris paribus* be applicable to both affirmative and negated propositions.

(3) The quantitative conditions of the occurrence of double CFs were tested. The choice of conjunctions as the most probable hypothesis here always involved double CFs. For all Pattern types there was a particular cell (e.g., “*A* & *B*” in 0° rotations) that had the highest frequency relative to the other three basic subclasses of a 2 × 2 contingency table. In Pattern Type 1 we obtained dominant double CFs corresponding to this cell. For Pattern Type 2, this prediction was corroborated as well, although $P_E(B \& F) < 0.5$. For Pattern Types 3 and 4, the most common selection did not correspond to conjunction *B* & *F* (or their rotations), even though this cell had a larger extensional probability than the dominant cell in Type 2. These results contradict the hypothesis that participants were only interested in the cell with the highest probability. This also rules out heuristics concerned with the majority of cases (which in other contexts may be crucial, see Schurz, 2005).

In summary, the results of Experiment 1 provide a first corroboration for BL and the predicted conditions under which double CFs occur. Extensional probabilities are obviously incoherent with the found systematic occurrence of double CFs. As will be argued in the General Discussion, no other theory concerning the CF has predicted double CFs under the qualitative and quantitative conditions explored here and none of the theories easily lends itself to explain the data.

6.2. Experiment 2: differential sample-size effects

Experiments 2a and 2b are mainly concerned with differential sample-size effects. Formalizing subjective probabilities, BL integrates the reliability of a sample and the probability of a logical hypothesis into a single measure. BL predicts that sample size should matter and that there may be different sample-size effects for different connectives.

Qualitatively, the studies again used strict and simple conditions: particularly, frequency information, no distractor hypotheses, and clear set inclusions. Quantitatively, they were concerned with varied probabilities (AND-patterns vs. A-patterns) and different sample sizes (low vs. high sample-size conditions). As a dependent variable, participants for each task had to select the hypothesis that appeared to be most probable. Additionally, participants had the alternative option to state that no hypothesis was the most probable. We predicted this option would be selected if the pattern probabilities of two hypotheses were nearly equal.

Fig. 2 shows the logical pattern probabilities calculated by BL for the four frequency patterns shown in Experiment 2a, assuming flat priors. For each pattern, two graphs are shown, illustrating the resultant probabilities for each investigated hypothesis: “*A* & *B*” (bank teller and feminist); “*A*”; and “*B*”. The left graph shows the resulting posterior probabilities (ordinate) for the hypotheses and

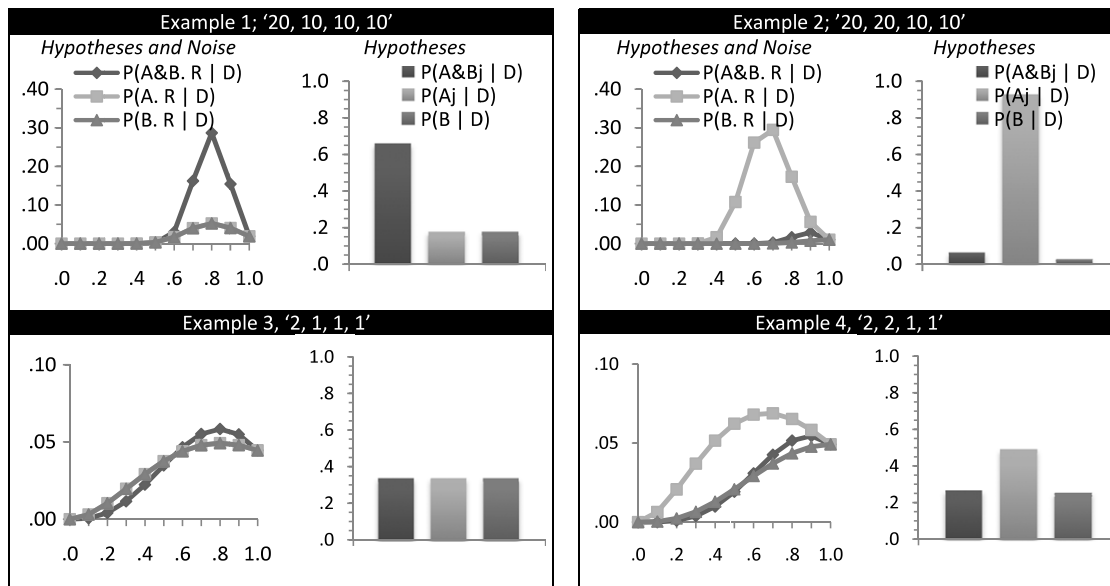


Fig. 2. Graphs of the resultant pattern probabilities $P_H(A \wedge B)$, $P_H(A)$, and $P_H(B)$, for uncertainty levels r between 0.0 and 1.0 ($P_H(A \circ_i B.r | D)$) (left graphs) and summing-up over all levels ($P_H(H_i | D)$) (right graphs), given the observed frequencies.

given noise levels r (abscissa) (Step 6 of the model), calculated from the given data. The graph on the right shows the posteriors of the hypotheses, integrating over all noise levels (Step 7). (Step 8 of the model, using log-odds, would reveal analogous predictions.)

In Fig. 2, Examples 1 and 2 show predictions for the large-sample-size conditions with $x_a = 20$, $x_b = 10$, $x_c = 10$, $x_d = 10$; and with $x_a = 20$, $x_b = 20$, $x_c = 10$, and $x_d = 10$. Assuming that many participants would interpret probability as pattern probabilities, BL predicted increased AND-selections for Example 1 (although $P_E(A \wedge B) < 0.5$) and increased A-selections (bank teller) in Example 2. Since both are used—the A and the B hypotheses along with the AND-hypothesis and no other filler items, a selection of AND-hypotheses extensionally involves committing double CFs. Examples 3 and 4 show the pattern probabilities for small sample-size conditions. The extensional probabilities are identical to the high sample-size conditions. Interestingly, the graphs demonstrate that the dominant hypotheses show a greater reduction in the small sample-size AND-condition (Example 3) than in the small sample-size A-condition (Example 4). They differ by only a single case, and their extensional probabilities are identical to the corresponding large set-size conditions (Examples 1 and 2). Hence a direct application of extensional probability should exhibit no sample-size effects. Alternatively, using some extensional sample statistics may predict an increase of “no preference”-choices in both low sample-size conditions, but no differential sample-size effects.

It should be noted that the outlined predictions of BL should be robust against (plausible) deviations from the assumption of a flat prior probability distribution for the noise levels. If the prior probability distribution were skewed, favoring either low or high noise levels, this “weighting factor” (according to BL) should not substantially alter judgment the most probable hypothesis unless this tendency is extreme. In Examples 2 and 4, the hypothesis A is always dominant over the other two hypotheses, independent of the assumed noise level. In Examples 1 and 3, the probability-ranks of the dominant hypotheses switch, even though this is not visible in Example 1. In Example 1, each individual conjunct has in fact a slightly higher likelihood than the conjunction for low noise levels, but for all other noise levels the likelihood of the conjunction is clearly higher. A prior-probability only moderately favoring low over high noise levels would not necessitate different predictions, since the higher noise levels are made probable by

the data (given all three possible logical hypotheses); hence the model's posterior of the hypotheses will almost certainly be based not on very low noise levels. Although the graph for Example 3 exhibits a switching of the rank-orders with slightly more clarity, the probabilities rarely differ substantially at a given noise level. Hence, it is only the presence of a strong prior belief in a quite low noise level (e.g., $P(r = 0.2) = 0.999$) that may overcome the impact of the likelihood function. In the studies, however, participants had no reason to believe strongly in such a prior. Hence the data-vectors employed allowed BL to make fairly robust predictions for the sample-size effects investigated.

6.2.1. Experiment 2a—differential sample-size effects in simultaneously presented tasks

Method: participants, materials, and procedure.

Ninety-six students from the University of Göttingen participated in two successive tasks concerned with the behaviour of graduates of different schools. Participants volunteered and received the same gratification as in Experiment 1. One participant was excluded for formal reasons.

The participants received a booklet (written in German) with paper-and-pencil tasks. The task-sheets were preceded by an introduction providing general information and an outline of the task's basic idea. The tasks were presented on successive pages, controlling for task order. Participants could observe four different frequencies corresponding to four schools: Linda (high frequency with an AND-pattern, i.e. High-AND-condition); Maria (High-A-condition); Sara (Low-AND-condition); and Nina (Low-A-condition). Participants were randomly assigned to one of eight presentation orders: (1) Linda-Sara; (2) Linda-Nina; (3) Maria-Sara; (4) Maria-Nina; (5) Sara-Linda; (6) Nina-Linda; (7) Sara-Maria; or (8) Nina-Maria. For each school, participants were presented with a 2×2 contingency table, representing frequency information about a sample of the schools' graduates (Table 4(a)). On the side of the table were labels: “bank tellers”, “not bank tellers”, “active feminists”, and “not active feminists”.

It was suggested to the participants that they consider an overall evaluation of each school. This received greater stress than it had in Experiment 1, in order to encourage pattern-answers rather than a response of “no preference” for participants in doubt. Furthermore, it was explicitly stated that a hypothesis was sought

Table 4
Observed frequency and percentages and numbers of selected hypothesis found for different schools in Experiment 2a.

Conditions	(a) Observed frequencies				(b) Percentage and number of choosing a hypothesis to be most probable				
	$A \wedge B$	$A \wedge \neg B$	$\neg A \wedge B$	$\neg A \wedge \neg B$	A	B	AND	?	n
'Linda School', High AND	20	10	10	10	17% 8	12% 6	40% 19	31% 15	48
'Maria School', High A	20	20	10	10	68% 32	6% 3	0% 0	26% 12	47
'Sara School', Low AND	2	1	1	1	14% 7	9% 4	13% 6	64% 30	47
'Nina School', Low A	2	2	1	1	54% 26	6% 3	6% 3	33% 16	48

Note. In Table 4(b) the predicted cells are darkened.

Table 5
Observed frequency and percentages and numbers of selected hypothesis found for different schools in Study 2b.

Conditions	(a) Observed frequencies				(b) Percentage and number of choosing a hypothesis to be most probable				
	$A \wedge B$	$A \wedge \neg B$	$\neg A \wedge B$	$\neg A \wedge \neg B$	A	B	AND	?	n
'Linda School', High AND	102	51	52	50	17% 8	17% 8	42% 20	25% 12	48
'Maria School', High A	102	100	50	52	67% 32	6% 3	15% 7	13% 6	48
'Sara School', Low AND	2	1	1	1	10% 5	6% 3	19% 9	65% 31	48
'Nina School', Low A	2	2	1	1	52% 25	4% 2	6% 3	38% 8	48

Note. In Table 5(b) the predicted cells are darkened.

that would describe the situation most adequately. Participants were informed that no hypothesis was strictly true. Based on the observed sample, and answering intuitively, participants were to place a checkmark next to “the hypothesis with the highest probability”.

The four hypotheses read as follows: “Today the girls of the Linda [Maria, etc.] school are generally [...]” (“Die Mädchen in der Linda Schule sind heute in der Regel”); (1) A: “bank tellers, whether they are feminists or not”. (Bankangestellte, egal ob sie aktive Feministinnen sind oder nicht”); (2) B: “active in the feminist movement, whether they are bank tellers or not”; (3) AND: “bank tellers who are active feminists” (“Bankangestellte, die”); or (4)? (the designation for “no preference”); that is, “with reference to the data, no single hypothesis is really better supported than any other”. There were no additional distractor hypotheses.

Results of Experiment 2a.

For the Linda-school, the AND-choice was selected more than twice as often as the A- and B-choices respectively (58%, not counting the “no preference” option). Notably, a relevant portion of participants chose the supplementary “no preference” option, presumably due to conflicting cues (some favoring pattern probabilities; some, standard extensional probabilities). Regarding presentation-order, the overall number of choices matching the predictions remained almost constant (62% in the first position, 67% in the second, with no significant difference ($\chi^2(1, n = 190) = 0.75, p = 0.45$)). Thus the data was collapsed over the counterbalancing variable Presentation order.

Statistical analysis corroborated that the portion of AND-choices – here always involving double CFs – was larger in the High-AND-condition than in all other conditions, including Low-AND (Linda vs. Sara, $\chi^2(1, n = 95) = 8.81, p < 0.01$). Additionally, the results confirmed that in the two low sample-size conditions, the AND-selections were not significantly different (Sara vs. Nina, exact Fisher test, $p = 0.32$). Interestingly, and as a contrast, more A-selections were found in the Low-A-condition than in the Low-AND-condition (Nina vs. Sara, $\chi^2(1, n = 95) = 16.16, p < 0.001$). As expected, the portion of A-selections was larger in the High-A-condition than in the High-AND-condition (Maria vs. Linda, $\chi^2(1, n = 95) = 25.76, p < 0.001$). Consistent with the predictions, the portion of A-selections varied neither between AND-conditions (Linda vs. Sara, $\chi^2(1, n = 95) = 0.06, p = 0.81$) nor between A-conditions (Maria vs. Nina,

$\chi^2(1, n = 95) = 1.93, p = 0.16$). Finally, the analysis corroborated a smaller increase of “no preference”-selections in the low sample-size conditions (relative to the high sample-size conditions) for the A-condition (Maria vs. Nina, $\chi^2(1, n = 95) = 0.70, p = 0.40$) than for the AND-conditions (Linda vs. Sara, $\chi^2(1, n = 95) = 10.11, p < 0.01$).

6.2.2. Experiment 2b

Replication of differential sample-size effects in simultaneously presented tasks.

The goal of Experiment 2b was to replicate the differential sample-size effects found in Experiment 2a, while ruling out possible objections. Experiment 2b resembled Experiment 2a, but differed in four main respects: (1) The tasks (the two schools) were presented simultaneously rather than successively, and on a single page. (2) Different frequency patterns were used for the high sample-size conditions (see Table 5(a)). In Experiment 2a the corresponding high and low frequency conditions employed identical probabilities. Since $P_E(A) = P_E(B)$ held in the Low-AND-condition, it needed to do so as well in the High-AND-condition. In 2b, however, contingency-tables were used (with $P_E(A) < P_E(B)$), precluding effects due to identical extensional probabilities. (3) In Experiment 2b, the dominant cell frequency in the High-A-condition corresponded to the AND-hypothesis, so that some theories of the CF would predict AND-selections here as well. (4) Finally, the sample size in the high frequency conditions was five times that seen in Experiment 2a, but according to BL, this should have no major effect. Thus the model's outputs closely corresponded to those depicted in Fig. 2, permitting the same predictions in Experiment 2b as in Experiment 2a.

Method: participants, materials, and procedure.

Ninety-six students of the University of Göttingen voluntarily participated in the experiment (cf. Experiment 1). Participants were concerned with two schools, presented on a single page, and told to choose in each case the hypothesis that was the “most probable” and “closest to the truth”. Their focus was assessing each school as a whole. A contingency table was provided, with frequency information on a sample of students from each school (see Table 5(a)). Participants were randomly assigned to serial orders, as in Experiment 2a, and were to check off the most

Table 6
Frequencies of observed cases in the different conditions of Experiment 3.

Sample sizes	Schools	$A \wedge B$	$A \wedge \neg B$	$\neg A \wedge \neg B$	$\neg A \wedge B$	Sum
Small	Linda, AND	18	5	6	7	36
	Maria, A	18	15	1	2	36
	Johanna, B	13	5	11	7	36
Medium	Linda, AND	39	11	12	13	75
	Maria, A	39	31	2	3	75
	Johanna, B	26	11	25	13	75
Large	Linda, AND	49	14	16	17	96
	Maria, A	49	40	3	4	96
	Johanna, B	33	14	32	17	96

probable option for each school. The hypotheses were presented along with a contingency table, formulated as in Experiment 2a.

Results.

Table 5(b) summarizes, by school, the number of participants along with the percentage who chose either a particular “most probable” hypothesis or the “no preference” option. For all of the conditions, the mode of the participants’ selections matched the BL predictions. Again, the task positions led to similar results for each school type, but presenting a school in the second position now significantly decreased the number of predicted answers ($\chi^2(1, n = 192) = 10.08; p < 0.01$). Nevertheless, the number of CFs in the critical High-AND-conditions in the two serial orders was identical.

Overall, the portion of AND-selections was significantly larger in the High-AND-condition than in the High-A-condition (Linda vs. Maria, $\chi^2(1, n = 96) = 8.71, p < 0.01$), with 56% double CFs in the High-AND-condition (without ‘no preference’ selections). The portion of A-choices was larger in the High-A than in the High-AND-condition (Linda vs. Maria, $\chi^2(1, n = 96) = 24.69, p < 0.001$). With respect to the predicted differential sample-size effect, the portion of AND-selections in the High-AND-condition was reduced in the Low-AND-condition (Sara vs. Linda, $\chi^2(1, n = 96) = 5.98, p < 0.05$); but, as predicted, the A-selections were not significantly reduced in the Low-A condition (Nina vs. Maria, $\chi^2(1, n = 96) = 2.12, p = 0.15$). Finally, there were more A-choices in Low-A than in Low-AND (Nina vs. Sara, $\chi^2(1, n = 96) = 19.39, p < 0.001$), and more “no preference”-selections in the Low-AND- than in the Low-A-condition (Sara vs. Nina, $\chi^2(1, n = 96) = 7.04, p < 0.01$).

6.2.3. Discussion of Experiments 2a and 2b

The results of the Experiments 2a and 2b corroborate the differential sample-size effects predicted by BL, and they replicate double-CF effects. As expected, similar results were obtained when the data were presented either successively (Experiment 2a) or simultaneously (Experiment 2b). Additionally, different frequencies were used. Finally, the mode of answers always corresponded to the BL prediction, depending systematically on frequency patterns as well as sample sizes.

The results were obtained despite the strict qualitative test conditions normally expected to favor an extensional interpretation of probability (e.g., Kahneman & Frederick, 2002, 2005). The possibility should be noted, however, that the relatively high number of “no preference”-selections over all conditions may reflect the uncertainty of task-interpretation (arguably due to conflicting cues).

Both studies’ results confirm differential sample-size effects, resulting in a stronger impact of different sample sizes (albeit same extensional probabilities) for the AND-condition than for the A-condition. Other accounts of the CF cannot explain this finding (see General Discussion). According to standard extensional probability, there should be no double CFs and since these probabilities remained identical (or almost) across varied sample-sizes one cannot predict these effects by any measure making

direct use of extensional probabilities. Furthermore, the unbiased estimator for extensional population probabilities remains the same, independent of sample size. Even if probability were interpreted in the sense of a statistical test, still referring to extensional probabilities, the uncertainty would be high for both low-frequency conditions, but with no differential effect.

6.3. Experiment 3: pattern sensitivity and rating scales

The goal of Experiment 3 is first and foremost to show whether probability judgments are pattern-sensitive, as predicted by BL. A second goal is to test that here varying sample sizes should have no (or only a weak) effect (cf. Study 2a and 2b). Finally, Experiment 3 qualitatively tests whether double CFs and single CFs can be obtained even if a rating format is added to the extensional cues.

(1) BL predicts internal and external pattern sensitivity. That is, BL predicts different ratings for $P_p(X)$ values, with extensional probabilities – $P_E(X)$ – remaining identical, due to changed distributions within or outside of a set. In the tasks, participants should provide probability ratings from samples of graduates, again differing in frequency patterns and sample sizes (see Table 6, cf. Fig. 3). External pattern sensitivity is tested in the Linda and Maria schools, where the overall set sizes and $P_E(A \wedge B)$ – hence $f(A \wedge B)$ – are held constant. Only the distribution – not the overall portion – of disconfirmatory cases outside of set $A \wedge B$ varied. Nevertheless, BL predicts that $P_p(A \wedge B)$ should be higher for the AND-pattern schools than for the respective A-pattern schools. Internal pattern-sensitivity effects should be caused by the distribution of cases within a set (here, “B”), again keeping the number of confirmative and disconfirmative cases constant. For the AND-pattern and B-pattern schools, $P_E(B)$ is identical, although for the latter $P_p(B)$ is higher (cf. Fig. 3). Significantly, pattern sensitivity has not been explicitly predicted by any other theory of the CF.

(2) Previously it has been shown that the use of rating formats instead of ranking formats substantially reduces the occurrence of CFs (Erdfelder, Bröder, & Brandt, 1998; Hertwig & Chase, 1998; Hertwig & Gigerenzer, 1999; Sloman et al., 2003; Wedell & Moro, 2008). Even Hertwig and Gigerenzer (1999) attributed roughly half the difference between single-case and frequency-conditions (cf. Fiedler, 1988) to the use of a rating-format. Experiment 3 we use natural frequencies, clear set inclusion, strict formulations and a rating format. According to BL, however, even this combination of extensional cues should be insufficient to suppress both CFs and double CFs.

(3) Whereas Experiments 2a and 2b each confirmed differential sample-size effects for rather low sample sizes, BL here predicts no substantial influence from sample-sizes, since confidence in the most probable hypothesis would increase only marginally with the higher sizes.

Method.

One hundred and eight students from the University of Göttingen participated in this experiment (cf. Experiment 1).

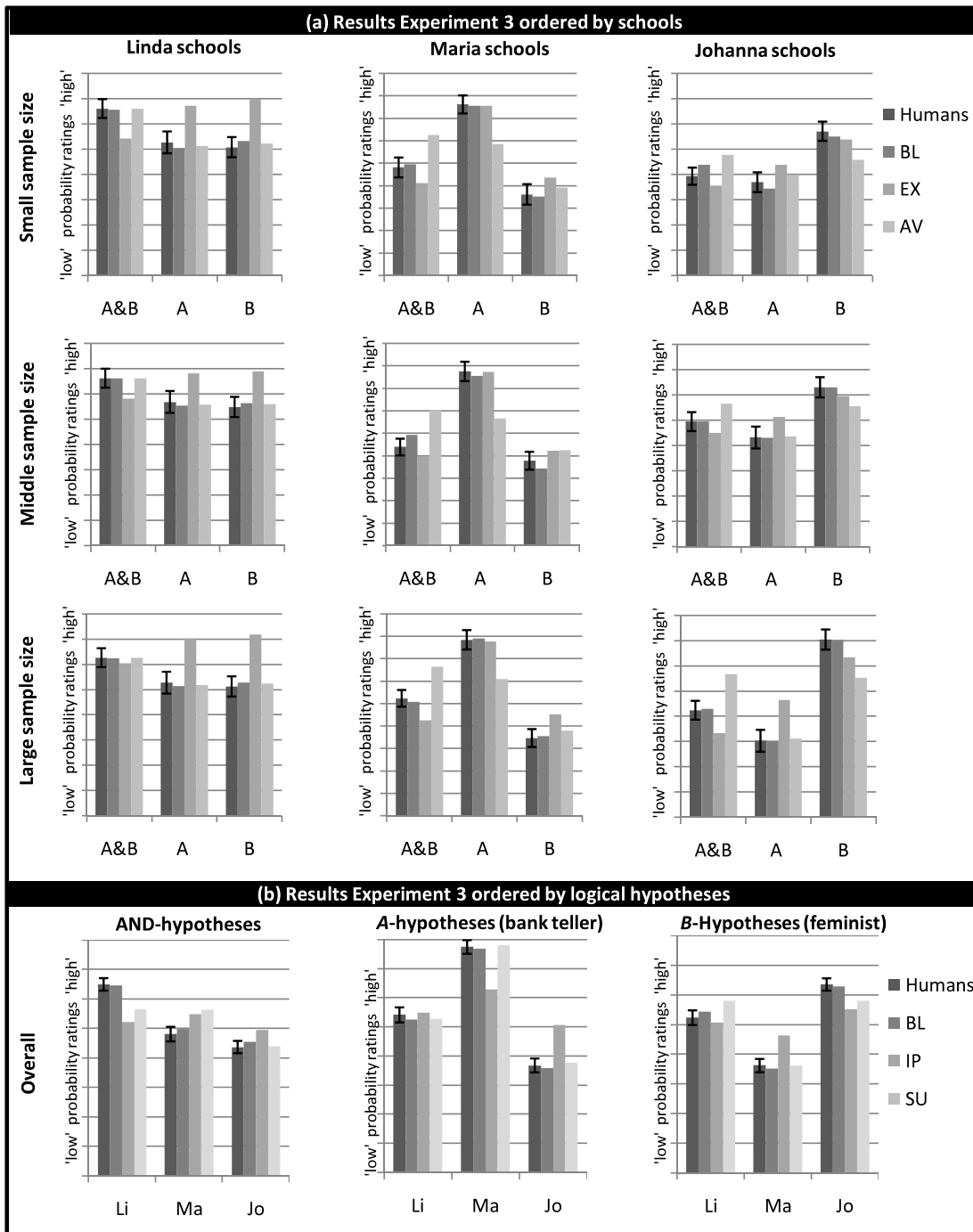


Fig. 3. (a) Actual rankings (on a scale with 11 steps) and data-based model predictions of Bayesian logic (BL), extensional probability (EX), and averaging heuristic (AV) in Study 3 for the three school types (Linda, Johanna, and Maria) and three sample sizes, and the three hypotheses. (b) Rankings and data-based model predictions of Bayesian logic (BL), inverse probability (IP), and support (SU) in different schools, ordered by connectives (here averaged over sample sizes). Error bars indicate standard errors.

Participants were to offer probability estimations based on rating scales for three schools (labeled Linda, Johanna and Maria). As in the previous studies, the samples for each school were presented in contingency tables (again, fulfilling the inequalities $P_E(A \wedge B) > P_E(A \wedge \neg B)$ and $P_E(A \wedge B) > P_E(\neg A \wedge B)$).

In this experiment, the schools were presented simultaneously. Instead of selecting a hypothesis, however, participants were to give three probability rankings for each school displaying the hypotheses *bank teller* (A), *feminist* (B), and *bank teller and feminist* (A & B). Table 6 lists the frequencies shown for each school. Three sample sizes ($N = 36, 75, 96$) were used, with analogous frequency patterns: Linda-school (AND-prediction),

Maria-school (A-prediction), and Johanna-school (B-prediction). The three schools shown to the participants each always had different sample sizes, in order to prevent direct comparisons. This resulted in six possible combinations of schools (L, M, J) and sample-sizes (S, M, L): LS-MM-JL; LL-MS-JM; LM-ML-JM; LS-ML-JM; LM-MS-JL; and LL-MM-JS (see Table 6). Hence each of the nine schools was present in two configurations. Additionally, the presentation order of the schools was counterbalanced (six orders, resulting in 36 possible tasks). Equal numbers of participants were assigned randomly to each task combination. Comparisons within sample-sizes allowed for testing of external and internal pattern sensitivity.

The headline of the task read (in German): “Hypotheses Concerning the Overall Evaluation of the Different Schools”. Instructions began: “Please indicate, in the context of evaluating the schools, for each school, which of the following *hypotheses* is most probable”. To counterbalance the use of rating scales, which could suggest standard extensional probability, it was explicitly stated that the focus be on understanding of the *whole situation* rather than on particular probabilities.

Participants were to provide three ratings for each of the schools, on scales spanning eleven steps, with extremes labeled “low” or “high” probability. It was decided not to use absolute values, to preclude participants’ calculating probabilities (which would have suggested a learned extensional meaning). Instead, participants were asked to provide intuitive judgments.

The three hypotheses, read as follows. “Girls from the Linda [Maria, etc.] school generally become ...” “H1: *bank tellers*, regardless of whether they are feminists or not” (“... , egal ob sie Feministinnen sind oder nicht”); “H2: *active feminists*, whether they are bank tellers or not”; or “H3: *bank tellers and at the same time feminists*” (“Bankangestellte und gleichzeitig Feministinnen”).

Finally, a hypothetical boss (recipient) was posited, who would be notified of the “most probable” hypothesis chosen for each school.

Results.

Fig. 3(a) and (b) show the results and the modeled predictions of various theories of the CF. Fig. 3(a) shows the average probability estimations (and mean errors) for each hypothesis for the different schools and sample sizes. It is apparent that the sample sizes (low, medium, and high) had no strong overall impact on the data.⁸ The figure also presents the predictions of BL (averaged posterior odds, model Step 8). The use of relative judgments is particularly plausible when using unnumbered rating scales. Model-fits directly using the posteriors (Step 7) will also be reported. The figure also presents the predictions of extensional reasoning (EX).

For all approaches, a linear regression was performed between the absolute predictions of the models and the empirical average ratings, in order to calculate the data-based predictions that provide a maximum likelihood estimate, given each model. Only this linear “correction” accounts for scaling-effects (clearly relevant for a scale without absolute values).

Fig. 3(a) shows a close fit between BL and the data. EX qualitatively never predicts double CFs. The proportion of systematic variance explained by BL for an average school (partial systematic eta square, the squared correlation of the data with the model) was $\eta^2 = 0.94$, if one uses the posterior probabilities (Step 7), or $\eta^2 = 0.97$, if one uses the log-odds (Step 8). Although the difference between these versions is not large, the results suggest participants are sensitive to the ratios of pattern probabilities presented in the graphs, as otherwise it would be impossible to predict the additional single CFs in the Maria- and Johanna-schools. With this plausible auxiliary hypothesis, all effects followed the direction predicted by BL. The results clearly favored BL over EX, which

⁸ For our mixed factorial design we also conducted an analysis of variance for the probability judgments with Hypotheses (A, B, A & B) and Schools (Linda, Maria, Johanna) as within-subjects variables, and Sample size (six patterns as between-subjects variables). The analysis yielded only one highly significant effect for the predicted interaction of Hypotheses and Schools: $F(4, 284) = 62.3, p < 0.0001, MSE = 5.7$. No other two-way interaction became significant. A three-way interaction, however, of Hypotheses, Schools, and Sample-size ($F(20, 284) = 1.6, p < 0.05, MSE = 5.7$) turned out to be significant as well, but further analysis showed that this could be attributed to the Johanna school only. Moreover, the effect size was small.

Table 7

Comparisons of probability-rankings between schools (*t*-tests for dependent samples) in Study 3.

Comparison	Diff.	SE	<i>t</i> ^a	<i>p</i>
Linda-AND vs. Maria-AND	1.69	0.33	5.08	<0.001
Maria-B vs. Linda-B	1.12	0.31	3.51	<0.001

^a *df* = 106 or 107.

Table 8

Comparisons of different probability ratings of the same schools (*t*-tests for dependent samples) in Study 3.

Comparison	Diff.	SE	<i>t</i> ^a	<i>p</i>
Linda-AND vs. Linda-A	1.08	0.32	3.31	<0.01
Linda-AND vs. Linda-B	1.26	0.35	3.60	<0.001
Linda-B vs. Linda-A	−0.18	0.28	−0.64	0.517
Maria-A vs. Maria-AND	2.92	0.35	8.15	<0.001
Maria-AND vs. Maria-B	1.20	0.29	4.10	<0.001
Johanna-B vs. Johanna-AND	1.98	0.32	6.08	<0.001
Johanna-AND vs. Johanna-A	0.69	0.24	2.80	<0.01

^a *df* = 106 or 107.

had a much lower model fit ($\eta^2 = 0.05$). (Note that Fig. 3 and particularly 3b include predictions of other CF models that will be discussed later in the article.⁹)

With regard to pattern-sensitivity, two comparisons (collapsing the data over Sample size) are of particular interest. External pattern sensitivity is indicated in Table 7, showing that $P_{emp}(A \wedge B)$ was estimated to be significantly higher in the Linda- than in the Maria-school, even though $P_E(A \wedge B)$ was held constant (see Table 6). Likewise, with regard to internal pattern sensitivity, a significant difference was found between $P_{emp}(A)$ in the Linda- and Johanna-conditions (Table 7).

In addition, within School types the comparisons of average ratings were consistent with BL (Table 8). The results confirmed double CFs in the Linda-school condition and single CFs in the Maria- and Johanna-school conditions despite the rating-response format.

Discussion of Experiment 3

The results of Experiment 3 confirm both external and internal pattern sensitivity. Patterns of frequencies do matter, even if the overall number of confirming or disconfirming frequencies is kept constant. These results were obtained even though Experiment 3 used a strict formulation for the AND-hypothesis and various other extensional cues (such as transparent tasks, full natural frequencies, clear set inclusions, and, in particular, a rating-response format). Most important, the results show pattern-sensitivity effects involving CFs and double CFs as predicted by BL.

7. General discussion

Bayesian logic provides a novel formalization of the polysemous term *probability* (Hertwig & Gigerenzer, 1999; Teigen, 1994)

⁹ BL and EX are most adequately modeled by comparing the predictions for each hypothesis given a particular school (Fig. 3(a)). Correspondingly, the results in Fig. 3(a) present for BL and EX (and AV, the averaging heuristic, with $\eta^2 = 0.51$) the fitted (linear regression) probability ratings of the hypotheses for each school type. In contrast, the theories of inverse probability (IP) and support (SU_{diff}) require comparisons of schools on a page, given a particular hypothesis (Fig. 3(b)), since these theories focus on a comparison of observed cases (in different schools) given a hypothesis. Fig. 3(b) provides a summary, averaging over sample sizes and the two combinations in which each school type was shown. The average explained systematic variance (accounting for school combinations) was $\eta^2 = 0.06$ for IP, $\eta^2 = 0.51$ for AV, and $\eta^2 = 0.70$ for SU_{diff} . To provide a fair comparison, BL can be analyzed on the level of *each* school combination (implying a lower *n*), yielding $\eta^2 = 0.86$ (posteriors) or $\eta^2 = 0.85$ (posterior odds). Even with these measures BL had a higher model fit than all other investigated models—notably, support was the only other model to achieve a substantial model fit, but qualitatively does not predict pattern-sensitivity effects.

concerned with underlying noisy-logical explanations of observed situations, predicting the occurrence of frequency-based CFs. The reported studies corroborate novel predictions of BL, including quantitative conditions of double CFs (Experiment 1), differential sample-size effects (Experiments 2a and 2b), and pattern-sensitivity effects (Experiment 3). Under the explored conditions, probabilities are assessed in a way that is highly consistent with BL.

In the discussion it is argued that the previous qualitative and quantitative theories of the CF do not offer a satisfactory explanation for the findings presented here. Although these theories concentrated on explaining single-event CFs, some of them may account for frequency-based CFs as well. Finally the relationship between BL and the representativeness heuristic is discussed, with final comments on the rationality of BL and domain-specific rational models in general.

7.1. Can previous theories of the CF explain the empirical findings?

7.1.1. Previous qualitative accounts of the CF

Previous findings in the CF-debate suggest that several cues elicit an extensional understanding of the task (Barbey & Sloman, 2007; Kahneman & Frederick, 2002, 2005). This extensional interpretation has mostly been regarded to be the only correct one. Although the focus was not on the extensional cues (i.e., the *qualitative* conditions of CFs), the results show that even simultaneous use even of all the cues can allow for large proportions of double CFs.

Here CFs occurred without single-event narratives, using full frequency information, clear set inclusion, transparent tasks, within-subjects comparisons, clarified formulations, no distractor hypotheses, and ranking formats. This novel finding goes beyond some previous impressive documentations of CFs (e.g., Lagnado & Shanks, 2002; Sloman et al., 2003; Tentori et al., 2004). When the goal is to assess the probabilities of logical hypotheses in order to provide an account of “the whole picture”, it appears that people evaluate which of several noisy-logical hypotheses is most likely to have produced exactly these data (in line with BL), even if all extensional cues are in place. Although extensional cues may be necessary for securing extensional reasoning, the results show that they are not sufficient to elicit extensional answers.

(1) *Frequency information and frequency questions.* In previous studies the proportion of CFs has often been substantially reduced by using natural-frequency rather than single-event formats (e.g., Hertwig & Gigerenzer, 1999; Kahneman & Frederick, 2002; Reeves & Lockhart, 1993; Tversky & Kahneman, 1983; by contrast, see Sloman et al., 2003; Wedell & Moro, 2008). Fiedler (1988) was the first to show in detail that frequency judgments involving 100 imaginary persons fitting the Linda description significantly reduced the rate of CFs. Based on such findings, Gigerenzer (1991, 1994, 1996, 1998, 2000) eloquently argued, from a frequentist perspective, that no fallacy was present, as long as CFs did not occur with frequencies. The conjunction “fallacy” and other “frequency illusions” would tend to disappear if presented in a natural-frequency format (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995, 2007; Sedlmeier & Gigerenzer, 2001; Zhu & Gigerenzer, 2006; cf. even Kahneman & Frederick, 2002, 2005).

In experiments reported here, by contrast, strong CF and double-CF effects were found based on natural frequencies alone, even if represented in contingency tables with information concerning the conjunction itself. In past studies, CFs were sometimes found in frequency tasks as well, using between-subject designs (Kahneman & Tversky, 1996), complex tasks (Lagnado & Shanks, 2002; Nilsson, 2008), or ranking tasks (Sloman et al., 2003). Wedell and Moro (2008) even showed the occurrence of CFs with frequency-response formats but did not provide frequency information on the conjunction. The novel findings

obtained here show substantial proportions of double CFs with the simultaneous use of full frequency information, transparent tasks, and a rating-response format.

After accounting for plausible confounding variables (clear set inclusions, ranking vs. rating, and so on), it appears necessary to distinguish two aspects of frequency formats: (1) the use of frequency information; and (2) the use of frequency questions (see also Girotto & Gonzales, 2001).

Frequency questions, by definition, requests a dependent variable that quantifies set-extensions. Hence such questions should elicit extensional reasoning and deviations from this remain fallacious (Tentori et al., 2004; Wedell & Moro, 2008). Even BL – at least in its current formulation – does not provide a measure of “the number of cases coherent with a pattern”. In the case of frequency questions, therefore, there should only be CFs when using few extensional formats and many cues triggering pattern probabilities (e.g., emphasizing the overall situation).

Frequency information can cue an extensional task interpretation (see Hertwig & Gigerenzer, 1999, Study 4) either by directly eliciting mathematical-extensional thinking or by emphasizing a nested-set relation (e.g., Hoffrage, Gigerenzer, Krauss & Martignon, 2002; Sloman & Over, 2003). In contrast to a frequentist interpretation, BL holds that CFs are not limited to single-event tasks; rather, they should occur with frequency information as well, as long as enough cues are present to elicit an “intensional” pattern understanding of the task. The obtained results corroborate this BL-prediction.

(2) *Set inclusion.* The salience of the nested-set relation between a conjunction and its conjuncts appears to be another important cue facilitating an extensional solution (Agnoli & Krantz, 1989; Neace et al., 2008; Sloman & Over, 2003; Sloman et al., 2003; cf. Barbey & Sloman, 2007; Evans et al., 2000; Johnson-Laird et al., 1999; Over, 2004). Admittedly, clear nested-set relations may in fact prevent CFs based on a misunderstanding of sets. Such misunderstandings are problematic for pattern probabilities as well. The reported studies show, however, that clear nested sets need to be supplemented by an *extensional use* of a given nested-set relation. Here strong double-CF effects have been found, despite the use of transparent nested-set relations.

(3) *Linguistic factors.* Subtle linguistic and pragmatic aspects of the task may exculpate participants from having committed CFs. In ordinary language, understanding of an affirmation “A” (e.g. active feminist) and a conjunction “A and B” (active feminist and bank teller) may differ from ideal logical usage (Hilton, 1995; Messer & Griggs, 1993; cf. Grice, 1975). That is, the affirmation may be understood as $A \wedge \text{not-}B$ (Agnoli & Krantz, 1989; Messer & Griggs, 1993; Morier & Borgida, 1984; Tentori et al., 2004), and the conjunction as $A \vee B, B|A$, or $A \rightarrow B$ (Hertwig et al., 2008; Mellers et al., 2001). Hence clarified formulations were used, such as “A, whether B or non-B” and “A, and at the same time B”. Experiment 1 additionally controlled for misunderstandings. In sum, although some CFs may previously have been due to linguistic misunderstandings, the systematic findings obtained here cannot be explained in this way.

(4) *Rating, ranking, or selection.* It is well-documented that a ranking-response format leads to more CFs than a rating-response format (Erdfelder et al., 1998; Hertwig & Chase, 1998; Neace et al., 2008; Sloman & Over, 2003; Sloman et al., 2003; Tversky & Kahneman, 1983, Exp. 4; Wedell & Moro, 2008; see also Hertwig & Gigerenzer, 1999). Only in relatively intransparent tasks was it found that rating formats did not substantially reduce the rate of CFs (Lagnado & Shanks, 2002). The current results show that a rating-response format is insufficient to eliminate the tendency to commit CFs, even in transparent tasks, if participant’s basic aim is to characterize an overall situation (see Experiment 3).

In sum, the experiments used several extensional cues simultaneously, but nonetheless yielded the systematic occurrence of CFs

and double CFs. Traditionally discussed extensional cues may be necessary to secure correct extensional reasoning, but the novel results show that they are by no means sufficient to elicit extensional reasoning, even if all are used simultaneously.

7.1.2. Quantitative theories of the CF

Some quantitative theories of the CF may appear applicable to the reported frequency-based tasks as well. Although previous quantitative accounts can explain some results, none can explain the overall evidence—quantitative conditions of double CF (Experiment 1), differential sample-size effects (Experiments 2a and 2b), and the effects of internal and external pattern-sensitivity (Experiment 3).

(1) *Tversky and Koehler's (1994) support theory* (which differs from subsequent theories of “support” or “confirmation”) concerns an unpacking process for varying numbers of subclasses. Since context stories and hypotheses were kept identical here, the theory would predict an equal number of CFs for all situations and cannot explain the CF-patterns found (cf. Crupi et al., 2008).

(2) *Inverse probability*. An influential explanation of traditional Linda-tasks posits that people may confuse posterior probabilities (such as $P(A \wedge B|L)$) with inverse probabilities ($P(L|A \wedge B)$) (Bar-Hillel, 1991; Fisk, 1996; Fisk & Slattery, 2005; Hertwig & Chase, 1998; Shafir, Smith, & Osherson, 1990; Wolford, 1991; Wolford, Taylor, & Beck, 1990). In principle, this might account for frequency-based CFs as well. With regard to the reported findings, however, there are three problems with this extensional approach. First, it cannot explain the double CFs found in Study 1. For in order to explain traditional Linda-tasks it has been assumed that the conjunctive subset “bank teller and feminist” is improbable at the outset; but if this is assumed, one would have to predict double CFs in all conditions, which is contrary to our findings. If, alternatively, the predictions were based on the observed cases (in the other schools), the contrastive character of this measure suggested effects based on presentation-order (which was not confirmed either). Second, the approach cannot explain differential sample-size effects (Studies 2a and 2b), since it takes probabilities as inputs rather than samples. Third, it cannot account for pattern-sensitivity effects (Study 3). It should be noted, finally, that inverse pattern probability might provide a related explanatory candidate.

(3) The *averaging theory of CFs* (Fantino, Kulik, Stolarz-Fantino, & Wright, 1997) posits that people may (falsely) average $P(A)$ and $P(B)$ to arrive at an estimate for $P(A \wedge B)$, sometimes resulting in CFs (Wedell & Moro, 2008; see in contrast Thüring & Jungermann, 1990). In contexts with no transparent information on the conjunction, some combination rule is needed in order to estimate this probability (Gavanski & Roskos-Ewoldsen, 1991; Nilsson, 2008; Waldmann, 2007). But it does not appear reasonable to apply averaging to the transparent frequency-table tasks investigated here, with direct evidence on the conjunction. Moreover, averaging would explain neither double CFs nor sample-size effects. Alternatively, one might propose an averaging heuristic, interpreting $P(A)$ as the average probability of its composing subsets (e.g., $P_{Av}(A) = (P_E(A \wedge B) + P_E(A \wedge \neg B))/2$). Such an account could predict double CFs but would have predicted them falsely when $P_E(\neg A \wedge B) < P_E(A \wedge \neg B) > P_E(A \wedge B)$ (i.e., in almost all of our studies, cf. Experiment 3). The theory of signed summation by Yates and Carlson (1986) would lead to similar predictions.

(4) *Rescaling*. The recently proposed theory of rescaling (Costello, 2005) modifies standard inductive logic (using standard conjunctive functions such as the product rule) by using a “probability” scale that goes beyond 1. Rescaling could explain some of the reported findings because it can explain CFs and can account in principle for double CFs. Nonetheless, any plausible setting of parameters in Experiment 1 necessitates the prediction that double

CFs will dominate either in all pattern types or in none. Additionally, rescaling — taking probabilities as inputs — cannot account for differential sample-size effects (Experiments 2a and 2b).

(5) *Support or confirmation*. The increase of extensional probability is another interesting candidate for explaining traditional CFs (Bonini, Tentori, & Osherson, 2004; Lagnado & Shanks, 2002; Sides et al., 2002). This assumes that subjects would substitute the target probability measure — $P_E(B \wedge F|D)$ — by a support measure (such as $P_E(B \wedge F|D) - P_E(B \wedge F)$ (SU_{Diff}) or $P_E(B \wedge F|D)/P_E(B \wedge F)$). Although it appears that adherents of this theory have not systematically investigated double CFs, support can explain them as well. Nevertheless, support cannot explain the quantitative conditions under which double CFs were found in Experiment 1. If predictions were derived based on the observed data, an effect of the different presentation orders would presumably have to be predicted. This (false) prediction can only be ruled out if in each task equal starting probabilities for all cells — $P_E(A \wedge B)$, $P_E(A \wedge \text{not-}B)$ etc. — are assumed, which came close to abandoning the core of the theory, concerned with differences of probabilities rather than absolute values. In any event, this would make the false prediction of double CFs for all pattern types. Moreover, support cannot explain the differential sample-size effects in Experiments 2a and 2b; and despite the second-best model fit in Experiment 3, it did not predict the pattern-sensitivity effects found. It is implausible that any of the measures of support (see Crupi et al., 2008) can mimic the predictions of BL corroborated here.

In summary, no other major theory of the CF can explain the qualitative and quantitative results obtained here. Although it is widely assumed that CFs constitute a homogeneous class of phenomena, several classes of CFs may in fact exist. The purpose of this paper has not been to rule out the existence of other classes of CFs, but to show that there is a particular class of frequency-based CFs that cannot be explained by traditional theories of the CF. It has been argued that this class may facilitate efficient communication about noisy-logical relations in an uncertain world.

7.1.3. Bayesian logic: a formalization of “representativeness” or an alternative rational model?

The representativeness heuristic was the first explanation for traditional CFs (Kahneman & Frederick, 2002, 2005; Kahneman & Tversky, 1973, 1972; Tversky & Kahneman, 1983, 1982). It should be asked, then, whether Bayesian logic is a more precise formalization of representativeness, or whether it differs from both concepts (i.e., extensional probability and representativeness).

Gigerenzer (1996, p. 592) has criticized the “one-word explanation” of representativeness for being “vague, undefined, and unspecified” (see also Gigerenzer, 1998; Nilsson, Juslin, & Olsson, 2008; Wolford, 1991). The representativeness heuristic has also been called empirically inadequate (Gavanski & Roskos-Ewoldsen, 1991; Gigerenzer, 1996; Nilsson, 2008; Nilsson et al., 2008). Kahneman and Frederick (2002, p. 73) and cf. Kahneman and Frederick (2005) defined their account of representativeness by clearly distinguishing two sub-processes: “(1) a prototype (a representative exemplar) is used to represent categories (e.g., bank teller) in the prediction task [Linda-task]; [and] (2) the probability that the individual belongs to a category is judged by the degree to which the individual resembles (is representative of) the category stereotype”. Their specification, however, does not provide a full mathematical model; moreover, Tversky and Kahneman (1983, p. 417) claimed that representativeness could not be computed as a formal function (such as product, sum, or weighted average of the scale values of its constituents). BL seems to provide a more precise formulation, at least with regard to frequency-based noisy-logical assessment of overall situations. Moreover, BL seems to possess traits reminiscent of representativeness, since pattern probabilities may be interpreted as a kind of similarity measure (although

differing from standard measures, as in Tversky, 1977). Furthermore, BL distinguishes between an extensional deliberate use and an intensional intuitive use, referring to pattern probabilities (De Neys, 2006; Kahneman & Frederick, 2002, 2005; Moutier & Houdé, 2003; Sloman & Over, 2003; cf. Osman, 2004). In these regards, BL seems to provide a mathematical specification of representativeness.

BL does clearly differ, however, in spirit, formalization and predictions, from the definition of representativeness by Kahneman and Frederick (2002, 2005). First, BL supposes that there is a rational class of CFs based on pattern probabilities of noisy-logical relations. In contrast, it has been essential to the idea of representativeness that it describes an irrational deviation from rational behavior (Kahneman & Frederick, 2002; Kahneman & Tversky, 1996; Tversky & Kahneman, 1983). Second, Kahneman and Frederick (2002, 2005) have argued that CFs should be strongly reduced under qualitative conditions, including clear set inclusions, within-subjects designs, and frequency information. Nor could this heuristic explain the quantitative conditions of the double CFs (Experiment 1). Since representativeness has mainly been concerned with qualitative tasks and single CFs, it is not clear how their account can apply to quantitative variations. Furthermore, Kahneman and Frederick have not allowed for conjunctive prototypes. It should be said, however, that representativeness might be extended, treating conjunctions as prototypic, if the conjunction represents the most frequent subclass. Yet even then, representativeness would predict double CFs when they were not obtained. Moreover, representativeness explains neither differential sample-size effects (Experiments 2a and 2b) nor the pattern-sensitivity effects (Experiment 3).

In conclusion, representativeness is either under-determined or unable to account for our results. Despite the similarities mentioned, BL makes different predictions and shows that CFs need not be fallacies at all, even if one is concerned with frequencies.

7.2. Is BL a rational model?

Whereas BL formalizes subjective pattern probabilities of logical connectives – $P_p(A \circ_l B|D)$ – by integrating over possible noise levels, extensional probabilities – $P_e(A \circ_l B|D)$ – are concerned with relative sizes of particular subsets only. BL, like several other theories of the CF, replaces the standard target measure of extensional probability estimations by another measure. In contrast to previous substitutions, however, BL is still concerned with probabilities of logical connectives, given the data. Thus, if subjects assess only $P(A \circ_l B|D)$, it is not inappropriate to apply pattern probability simply because experimenters apply another measure. If one is concerned with assessing an overall situation, BL provides a rational model of probability-judgment for noisy-logical relations; at least it is the most rational model currently at hand. BL hence allows for what one may paradoxically refer to as “rational CFs”.

BL is concerned with logically nested sets and only applies if corresponding misunderstandings are ruled out. For instance, “ $A \wedge B$ ” is understood to be nested in “ A ” (cf. Hertwig et al., 2008; Mellers et al., 2001). Nevertheless, logical pattern probabilities do violate the equation $P_p(A \wedge B|D) + P_p(A \wedge \text{non-}B|D) = P_p(A|D)$. Since the logical hypotheses are treated as alternatives, however, it appears that, in a different sense, noisy connectives are not nested. BL subscribes to additivity at the different level of alternative hypotheses. Within this framework, it should therefore be impossible to construct a Dutch-book argument against the use of pattern probabilities of noisy-logical relationships.

7.3. The rationality debate and domain-specific normative models

The confirmation of BL sheds light on the general rationality debate in psychology, particularly between Gigerenzer and colleagues on the one hand, and Kahneman, Tversky and colleagues on the other. Although it criticizes Gigerenzer’s frequentist approach, the mathematical formulation of inductive Bayesian logic was actually partly inspired by his quest to counter content-blind application of narrow domain-general norms and vague heuristics (Gigerenzer, 1991, 1996, 1998, 2000).

Nevertheless, the advocated approach asserts that Kahneman and Tversky (1996), see also Kahneman and Frederick (2005), with many others, before and after, were right to oppose “normative agnosticism”. Yet content-blind universalism must be supplemented by content-sensitive “local” norms of rationality. BL is a domain-specific account, applicable conditional on the satisfaction of the model’s preconditions in the actual world (“conditional normativity”).

Even if most model assumptions appear necessary, some might need to be modified in the future. The significant distinction of normative versus descriptive sides of reasoning – historically related to what has been called *rational vs. empirical psychology* – does not imply we need to claim omniscience regarding the applicability of a particular norm. Not only empirical knowledge but also knowledge about norms of rationality have undergone refinement in the past, and they will be refined in the future. Moreover, the dichotomy of normative and descriptive accounts need not imply an eternally unbridgeable abyss. Rather than being necessarily attributed to biased reasoning, a mismatch between a normative theory and actual reasoning-behavior may suggest that the normative model was inappropriate in the first place (see Chater & Oaksford, 2000). Finally, although explanations at the computational level, such as BL, are important, they do not preclude the investigation of simpler heuristics that may approximate the results of such rational models.

In conclusion, the advocated proposal pursues a third way between upholding the Enlightenment vision of universal norms of rationality, which in psychology led to an emphasis on the fallacious character of human thought (e.g., Kahneman and Tversky), and an evolutionary approach, postulating domain-specific adaptations to specific environments (e.g., Gigerenzer). Here an attempt has been made to forge ahead on this third way of domain-specific yet rational norms of thought.

Acknowledgments

Thanks to Anna Gast, Christina Botros, Deborah Wolff, Johanna Frisch, Kristin Reißer, and Mareike Methling, for running the studies. I would like to thank John K. Kruschke and anonymous reviewers for their valuable comments on a previous version of the manuscript. Additionally, I am grateful to Björn Meder, Johanna Frisch, Jonathan Nelson, Larry Fiddick, Martha Cunningham, and Ralf Mayrhofer for earlier comments, and to the Waldmann-Hagmayer lab at Göttingen for inspiration.

This work has been funded through a grant “Bayeslogik” by the *Deutsche Forschungsgemeinschaft* (DFG, Sy 111/1-2). The author is also member of the Courant Research Center, *Evolution of Social Behaviour* (Kellnerweg 6, D-37077 Göttingen; supported by the German Excellence Initiative).

Appendix A

See Fig. 4.

Overall Situation at the Linda School

You are interested in an **overall evaluation of this school and its graduates**. Do the graduates of this school generally become bank tellers, feminist or both at the same time? You are concerned with different hypotheses concerning the occurrence of these properties and how these properties are related at this school. In the given situations **none of the hypotheses is 100% correct**. It is your task, to tick the hypothesis that—based on the given sample—appears to hold at least most probable in this situation. We are interested in your intuitive judgment.

For each school please tick only one proposition that appears most probable to you. Only this information will be transferred to your boss.

Sample of graduate of the Linda School

	Feminist	No feminist
Bank teller	Bank teller & feminist 48	Bank teller & no feminist 6
No bank teller	No bank teller & feminist 2	No bank teller & no feminist 4

Most test conditions are easy. We are interested in your intuitive judgment.

- The graduates of the Linda School generally become **bank tellers and at the same time feminists**.
- The graduates of the Linda School generally become **bank tellers and at the same time no feminists**.
- The graduates of the Linda School generally become **no bank tellers and at the same time feminists**.
- The graduates of the Linda School generally become **no bank tellers and at the same time no feminists**.
- The graduates of the Linda School generally become **bank tellers**.
- The graduates of the Linda School generally become **no bank tellers**.
- The graduates of the Linda School generally become **feminists**.
- The graduates of the Linda School generally become **no feminists**.

Fig. 4. Instruction of a frequency-based probability judgment task in Experiment 1 (translated).

Appendix B. Log-linear analysis of Experiment 1

A parsimonious way to test whether rotation and presentation-positions had an impact on the results is to use log-linear analysis, simply treating the data as independent observations. Note, however, that this only accounts for a task's position, ignoring which particular test preceded which task. Such within-group comparisons are common, although they may limit the interpretability (Kennedy, 1992). Nonetheless, the analysis is interesting and permissible. In fact, only a small portion of comparisons is based on within-subject data, and each data-point was directly preceded by all other different patterns equally often. To increase the number of expected cases in each cell, we reduced the overall table, using only as factors Pattern types (Types 1 and 2 vs. Types 3 and 4) \times Rotations \times Positions (Positions 1 and 2 vs. 3 and 4) \times Possible answers (*B* & *F* vs. *B* vs. Rest). BL predicts that one exclusive interaction between Pattern-type and Answers should fit the data. Our analysis tested for the highest-order interaction term needed in the model, including all hierarchically lower effects, and yielded that neither 3-way nor 4-way terms are needed; only the 2-way interaction became significant ($k = 2$ with Pearson's $\chi^2 = 31.26$, $df = 17$, $p = 0.01$; $k = 3$ with $\chi^2 = 15.64$, $df = 17$, $p = 0.54$; and $k = 4$ with $\chi^2 = 5.47$, $df = 6$, $p = 0.48$). The automatic model-fitting (with delta = 0.5) yielded only the predicted second-order term (Pattern-type \times Answers), which fitted the data without significant deviation ($\chi^2 = 28.60$, $df = 42$, $p = 0.94$).

Although log-linear models are robust against moderate violations of the necessary number of expected cases (here 33% instead of 20% had under five cases, but there was only one zero cell), a further analysis was run, enlarging n per cell by distinguishing only two types of rotations. The results again yielded only the predicted second-order term (Pattern-type \times Answers), which again fit the data well ($\chi^2 = 12.12$, $df = 18$, $p = 0.84$).

References

- Adams, E. W. (1986). On the logic of high probability. *Journal of Philosophical Logic*, 1, 255–279.
- Agnoli, F., & Krantz, D. (1989). Suppressing natural heuristics by formal instruction: the case of the conjunction fallacy. *Cognitive Psychology*, 21, 515–550.
- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale: Erlbaum.
- Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: from ecological rationality to dual processes. *Behavioral and Brain Sciences*, 30, 241–297.
- Bar-Hillel, M. (1991). Commentary on Wolford, Taylor, and Beck: the conjunction fallacy. *Memory & Cognition*, 19, 412–414.
- Betsch, T., & Fiedler, K. (1999). Understanding conjunction effects: the role of implicit mental models. *European Journal of Social Psychology*, 29, 75–93.
- Bonini, N., Tentori, K., & Osherson, D. (2004). A different conjunction fallacy. *Mind & Language*, 19, 199–210.
- Bussemeyer, J. R., Wang, Z., & Townsend, J. T. (2006). Quantum dynamics of human decision-making. *Journal of Mathematical Psychology*, 50, 220–241.
- Carbonetto, P., Kisyński, J., de Freitas, N., & Poole, D. (2005). Nonparametric Bayesian logic. In *Proc. 21st UAI* (pp. 1–9).
- Carnap, R. (1952). *The continuum of inductive methods*. Chicago University Press.
- Chater, N., & Oaksford, M. (2000). The rational analysis of mind and behavior. *Synthese*, 122, 93–131.
- Chater, N., & Oaksford, M. (Eds.). (2008). *The probabilistic mind: prospects for Bayesian cognitive science*. Oxford University Press.
- Chater, N., Tenenbaum, J., & Yuille, A. (2006). Probabilistic models of cognition: conceptual foundations. *Trends in Cognitive Sciences*, 10, 287–291.
- Cheng, P. W. (1997). From covariation to causation: a causal power theory. *Psychological Review*, 104, 367–405.
- Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58, 1–73.
- Costello, F. J. (2005). A unified account of conjunction and disjunction fallacies in people's judgments of likelihood. In *Proceedings of the twenty-seventh annual conference of the cognitive science society* (pp. 494–499). Mahwah, NJ: Erlbaum.
- Crupi, V., Fitelson, B., & Tentori, K. (2008). Probability, confirmation, and the conjunction fallacy—theoretical note. *Thinking & Reasoning*, 14, 182–199.
- De Houwer, J., & Beckers, T. (2002). A review of recent developments in research and theories on human contingency learning. *The Quarterly Journal of Experimental Psychology, Section B*, 55, 289–310.
- De Neys, W. (2006). Automatic-heuristic and executive-analytic processing during reasoning: chronometric and dual-task considerations. *Quarterly Journal of Experimental Psychology*, 59, 1070–1100.
- De Raedt, L., & Kersting, K. (2003). Probabilistic logic learning. *SIGKDD Exploitations*, 5(1), 31–48.
- Erdfelder, E., Bröder, A., & Brandt, M. (1998). Zur bedeutung des antwortformats für die Häufigkeit von konjunktionsfehlern. The relevance of response format for the frequency of conjunction errors. In *Talk presented at the 40th TEAP in Marburg, Germany*.

- Evans, J. St. B. T., Handley, S. H., Perham, N., Over, D. E., & Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. *Cognition*, 77, 197–213.
- Evans, St. B. T., & Over, D. E. (2004). *If*. Oxford University Press.
- Fantino, E., Kulik, J., Stolarz-Fantino, St., & Wright, W. (1997). The conjunction fallacy: a test of averaging hypotheses. *Psychonomic Bulletin & Review*, 4, 96–101.
- Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research*, 50, 123–129.
- Fiedler, K. (2000). Beware of samples! a cognitive-ecological sampling approach to judgment biases. *Psychological Review*, 107, 659–676.
- Fisk, J. E. (1996). The conjunction effect: fallacy or Bayesian inference? *Organizational Behavior and Human Decision Processes*, 67, 76–90.
- Fisk, J. E., & Slattery, R. (2005). Reasoning about conjunctive probabilistic concepts in childhood. *Canadian Journal of Experimental Psychology*, 59, 168–178.
- Fitelson, B. (2006). Inductive logic. In J. Pfeifer, & S. Sarkar (Eds.), *Philosophy of science: an encyclopedia*. London: Routledge.
- Frege, G. (1879). *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*. Halle: Verlag Louis Nebert.
- Gavanski, I., & Roskos-Ewoldsen, D. R. (1991). Representativeness and conjoint probability. *Journal of Personality and Social Psychology*, 61, 181–194.
- Gigerenzer, G. (1991). How to make cognitive illusions disappear: beyond 'heuristics and biases'. In W. Stroebe, & M. Hewstone (Eds.), *European review of social psychology*. Vol. 2 (pp. 83–115). Chichester: Wiley.
- Gigerenzer, G. (1994). Why the distinction between single-event covariations and frequencies is relevant for psychology and vice versa. In G. Wright, & P. Ayton (Eds.), *Subjective probability* (pp. 129–161). Chichester: Wiley.
- Gigerenzer, G. (1996). On narrow norms and vague heuristics: a reply to Kahneman and Tversky. *Psychological Review*, 103, 592–596.
- Gigerenzer, G. (1998). Ecological intelligence: an adaptation for frequencies. In D. D. Cummins, & C. Allen (Eds.), *The evolution of mind*. New York: Oxford University Press.
- Gigerenzer, G. (2000). *Adaptive thinking: rationality in the real world*. London: Oxford University Press.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review*, 102, 684–704.
- Gigerenzer, G., & Hoffrage, U. (2007). The role of representation in Bayesian reasoning: correcting common misconceptions. *Behavioral and Brain Sciences*, 30, 264–267.
- Gilovich, Th., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases. The psychology of intuitive judgement*. Cambridge University Press.
- Giroto, V., & Gonzales, M. (2001). Solving probabilistic and statistical problems: a matter of information structure and question form. *Cognition*, 78, 247–276.
- Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, 32, 108–154.
- Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. Morgan (Eds.), *Syntax and semantics: Vol. 3* (pp. 41–58). New York: Academic Press.
- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51, 334–384.
- Hagmayer, Y., Sloman, S. A., Lagnado, D. A., & Waldmann, M. R. (2007). Causal reasoning through intervention. In A. Gopnik, & L. Schulz (Eds.), *Causal learning: psychology, philosophy, and computation* (pp. 86–100). Oxford: Oxford University Press.
- Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: a Bayesian approach to reasoning fallacies. *Psychological Review*, 114, 704–732.
- Hájek, A. (2001). Probability, logic, and probability logic. In L. Goble (Ed.), *The blackwell companion to logic* (pp. 362–384). Oxford: Blackwell.
- Hájek, P. (2002). Fuzzy logic. In: Zalta, E.N. (Ed.), *The stanford encyclopedia of philosophy*. URL: <http://plato.stanford.edu/entries/logic-fuzzy/> (3rd September).
- Hattori, M. (2002). A quantitative model of optimal data selection in Wason's selection task. *Quarterly Journal of Experimental Psychology, Section A*, 55, 1241–1272.
- Hattori, M., & Oaksford, M. (2007). Adaptive non-interventional heuristics for covariation detection in causal induction: model comparison and rational analysis. *Cognitive Science*, 31, 765–814.
- Hempel, C. G. (1945). Studies in the logic of confirmation. *Mind*, 54, 1–25. 7–121.
- Hertwig, R., Benz, B., & Krauss, B. S. (2008). The conjunction fallacy and the many meanings of and. *Cognition*, 108, 740–753.
- Hertwig, R., & Chase, V. M. (1998). Many reasons or just one: how response mode affects reasoning in the conjunction problem. *Thinking & Reasoning*, 4, 319–352.
- Hertwig, R., & Gigerenzer, G. (1999). The 'conjunction fallacy' revisited. *Journal of Behavioral Decision Making*, 12, 275–305.
- Hilton, D. J. (1995). The social context of reasoning: conversational inference and rational judgment. *Psychological Bulletin*, 118, 248–271.
- Hintikka, J. (2004). A fallacious fallacy? *Synthese*, 140, 25–35.
- Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition*, 84, 343–352.
- Johnson-Laird, P. N., & Byrne, R. M. J. (2002). A theory of meaning, pragmatics, and inference. *Psychological Review*, 109, 646–678.
- Johnson-Laird, P. N., Legrenzi, P., Giroto, V., Legrenzi, S. M., & Caverni, J.-P. (1999). Naive probability: a mental model theory of extensional reasoning. *Psychological Review*, 106, 62–88.
- Kahneman, D., & Frederick, S. (2002). Representativeness revised: attribute substitution in intuitive judgment. In Th. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases. The psychology of intuitive judgement* (pp. 49–81). Cambridge University Press.
- Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. J. Holyoak, & R. G. Morris (Eds.), *The cambridge handbook of thinking & reasoning* (pp. 267–293). Cambridge University Press.
- Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, 103, 582–591.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237–251.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cognitive Psychology*, 3, 430–454.
- Kao, S.-F., & Wassermann, E. A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 1363–1386.
- Kennedy, J. J. (1992). *Analyzing qualitative data: log-linear analysis for behavioral research* (2nd ed.) New York: Praeger.
- Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer.
- Kruschke, J. K. (2008). Bayesian approaches to associative learning: from passive to active learning. *Learning & Behavior*, 36, 210–226.
- Lagnado, D. A., & Shanks, D. R. (2002). Probability judgment in hierarchical learning: a conflict between predictiveness and coherence. *Cognition*, 93, 81–112.
- Marr, D. (1982). *Vision: a computational approach*. San Francisco: Freeman & Co.
- McKenzie, C. R. M., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology*, 54, 33–61.
- Mellers, B. A., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, 12, 269–275.
- Messer, W. S., & Griggs, R. A. (1993). Another look at Linda. *Bulletin of the Psychonomic Society*, 31, 193–196.
- Milch, B., Marthi, B., Russell, S., Sontag, D., Ong, D. L., & Kolobov, A. (2006). BLOG: probabilistic models with unknown objects. *Dragstuhl Seminar Proceedings*, 2005, 1–6.
- Morier, D. M., & Borgida, E. (1984). The conjunction fallacy: a task specific phenomenon? *Personality and Social Psychology Bulletin*, 10, 243–253.
- Moutier, S., & Houdé, O. (2003). Judgement under uncertainty and conjunction fallacy inhibition training. *Thinking and Reasoning*, 9, 185–201.
- Neace, W. P., Michaud, St., Bolling, L., Deer, K., & Zecevic, L. (2008). Frequency formats, probability formats, or problem structure? A test of the nested-sets hypothesis in an extensional reasoning task. *Judgment and Decision Making*, 3, 140–152.
- Nelson, J. (2005). Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, 112, 979–999.
- Nickerson, R. S. (2004). *Cognition and chance. The psychology of probabilistic reasoning*. Mahwah: Lawrence Erlbaum Associates.
- Nickerson, R. S. (1996). Hempel's paradox and Wason's selection task: logical and psychological puzzles of confirmation. *Thinking and Reasoning*, 4, 231–248.
- Nilsson, H. (2008). Exploring the conjunction fallacy within a category learning framework. *Journal of Behavioral Decision Making*, 21, 471–490.
- Nilsson, H., Juslin, P., & Olsson, H. (2008). Exemplars in the mist: the cognitive substrate of the representativeness heuristic. *Scandinavian Journal of Psychology*, 49, 201–212.
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality. The probabilistic approach to human reasoning*. Oxford: Oxford University Press.
- Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101, 608–631.
- Oaksford, M., & Chater, N. (2003). Optimal data selection: revision, review, and reevaluation. *Psychological Bulletin & Review*, 10, 289–318.
- Oberauer, K., Weidenfeld, A., & Fischer, K. (2007). What makes us believe a conditional? The roles of covariation and causality. *Thinking and Reasoning*, 13, 340–369.
- Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, 11, 988–1010.
- Over, D. E. (2004). Naïve probability and its model theory. In V. Giroto, & P. N. Johnson-Laird (Eds.), *The shape of reason: essays in honour of Paolo Legrenzi* (pp. 139–160). Hove: Psychology Press.
- Over, D. E., Hadjichristidis, C., Evans, J. St. B. T., Handley, S. J., & Sloman, St. A. (2007). The probability of causal conditionals. *Cognitive Psychology*, 54, 62–97.
- Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press.
- Piaget, J., & Garcia, R. (1991). In Ph. M. Davidson, & J. Easley (Eds.), *Toward a logic of meanings*. Hillsdale: Erlbaum.
- Pineno, O., & Miller, R. R. (2007). Comparing associative, statistical, and inferential reasoning accounts of human contingency learning. *The Quarterly Journal of Experimental Psychology*, 60, 310–329.
- Popper, K. R. (2005). *Logik der Forschung* (11th ed.) Tübingen: Mohr-Siebeck. (Original work published 1934).
- 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.
- Reichenbach, H. (1935). Wahrscheinlichkeitslogik [probability logic]. *Erkenntnis*, 5, 37–43.
- Romeyn, J. W. (2005). Bayesian inductive logic. In *Inductive prediction from statistical hypotheses*. Doctoraat aan de Wijsbegeerte, Rijksuniversiteit Groningen.
- Schurz, G. (2001). Normische Gesetzhypothesen und die wissenschaftsphilosophische Bedeutung des nichtmonotonen Schliessens. *Journal for General Philosophy of Science*, 32, 65–107.

- Schurz, G. (2005). Non-monotonic reasoning from an evolutionary viewpoint: ontic, logical and cognitive foundations. *Synthese*, 146, 37–51.
- Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, 130, 380–400.
- Shafir, E., Smith, E., & Osherson, D. (1990). Typicality and reasoning fallacies. *Memory & Cognition*, 18, 229–239.
- Sides, A., Osherson, D., Bonini, N., & Viale, R. (2002). On the reality of the conjunction fallacy. *Memory & Cognition*, 30, 191–198.
- Skyrms, B. (1986). *Choice and chance. An introduction to inductive logic* (3rd ed.). Belmont, CA: Wadsworth Publishing Company.
- Sloman, S. A., & Lagnado, D. (2005). Do we “do”? *Cognitive Science*, 29, 5–39.
- Sloman, S. A., & Over, D. (2003). Probability judgement from the inside and out. In D. Over (Ed.), *Evolution and the psychology of thinking: the debate* (pp. 145–169). Hove, New York: Psychology Press.
- Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes*, 91, 296–309.
- Stalnaker, R. (1968). A theory of conditionals. *American Philosophical Quarterly Monograph Series*, 2, 98–112.
- Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgement. *Psychological Review*, 112, 881–911.
- Stolarz-Fantino, S., Fantino, E., & Kulik, J. (1996). The conjunction fallacy: differential incidence as a function of descriptive frames and educational context. *Contemporary Educational Psychology*, 21, 208–218.
- Teigen, K. H. (1994). Variants of subjective probabilities: concepts, norms, and biases. In G. Wright, & P. Ayton (Eds.), *Subjective probability* (pp. 211–238). Chichester: John Wiley.
- Tenenbaum, J., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24, 629–640.
- Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science*, 10, 309–318.
- Tentori, K., Bonini, N., & Osherson, D. (2004). The conjunction fallacy: a misunderstanding about conjunction? *Cognitive Science*, 28, 467–477.
- Thüring, M., & Jungermann, H. (1990). The conjunction fallacy: causality vs. event probability. *Journal of Behavioral Decision Making*, 3, 61–74.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327–352.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review*, 90, 293–315.
- 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 University Press.
- Tversky, A., & Koehler, D. J. (1994). Support theory: a nonextensional representation of subjective probability. *Psychological Review*, 101, 547–567.
- von Sydow, M. (2009). On a general Bayesian pattern logic of frequency-based logical inclusion fallacies. In *Proceedings of the thirty-first annual conference of the cognitive science society* (pp. 248–253). Austin, TX: Cognitive Science Society.
- von Sydow, M. (2007a). The Bayesian logic of the conjunction fallacy: effects of probabilities and frequencies in contingency tables. In *Proceedings of the twenty-ninth annual conference of the cognitive science society* (pp. 1611–1616). Austin, TX: Cognitive Science Society.
- von Sydow, M. (2008). The Bayesian logic of conjunction fallacies: probability rating tasks and pattern-sensitivity. In *Proceedings of the twenty-ninth annual conference of the cognitive science society* (pp. 1795–1800). Austin, TX: Cognitive Science Society.
- von Sydow, M. (2007b). Testing descriptive or prescriptive conditionals and differential effects of frequency information. In *Proceedings of the twenty-ninth annual conference of the cognitive science society* (pp. 1617–1622). Mahwah, NJ: Erlbaum.
- von Sydow, M. (2006). Towards a flexible Bayesian and deontic logic of testing descriptive and prescriptive rules. *Doctoral dissertation in psychology*, Georg-August-Universität Göttingen.
- von Sydow, M. (2004). Structural Bayesian models of conditionals. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the twenty-sixth annual conference of the cognitive science society* (pp. 1411–1416). Mahwah, NJ: Erlbaum.
- Vranas, P. B. M. (2004). Hempel's raven paradox. Lacuna in the standard Bayesian solution. *British Journal for the Philosophy of Science*, 55, 545–560.
- Waldmann, M. R. (2007). Combining versus analyzing multiple causes: how domain assumptions and task context affect integration rules. *Cognitive Science*, 31, 233–256.
- Wedell, D., H., & Moro, R. (2008). Testing boundary conditions for the conjunction fallacy: effects of response mode, conceptual focus, and problem type. *Cognition*, 107, 105–136.
- White, P. A. (2002). Causal attribution from covariation information: the evidential evaluation model. *European Journal of Social Psychology*, 32, 667–684.
- Wolford, G. (1991). The conjunction fallacy? A reply to Bar-Hillel. *Memory & Cognition*, 19, 415–417.
- Wolford, G., Taylor, H. A., & Beck, J. R. (1990). The conjunction fallacy? *Memory & Cognition*, 18, 47–53.
- Yates, J. F., & Carlson, B. W. (1986). Conjunction errors: evidence for multiple judgment procedures, including 'signed summation'. *Organizational Behavior and Human Decision Processes*, 37, 230–253.
- Zhu, L., & Gigerenzer, G. (2006). Children can solve Bayesian problems: the role of representation in mental computation. *Cognition*, 98, 287–293.