COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 35 (2011) 682–711
Copyright © 2011 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2011.01170.x

# Source Reliability and the Conjunction Fallacy

Andreas Jarvstad, Ulrike Hahn

*School of Psychology, Cardiff University*

Received 21 September 2009; received in revised form 2 November 2010; accepted 4 November 2010

**Abstract**

Information generally comes from less than fully reliable sources. Rationality, it seems, requires that one take source reliability into account when reasoning on the basis of such information. Recently, Bovens and Hartmann (2003) proposed an account of the conjunction fallacy based on this idea. They show that, when statements in conjunction fallacy scenarios are perceived as coming from such sources, probability theory *prescribes* that the ‘‘fallacy’’ be committed in certain situations. Here, the empirical validity of their model was assessed. The model predicts that statements added to standard conjunction problems will change the incidence of the fallacy. It also predicts that statements from reliable sources should yield an increase in fallacy rates (relative to unreliable sources). Neither the former (Experiment 1) nor the latter prediction (Experiment 3) was confirmed, although Experiment 2 showed that people can derive source reliability estimates from the likelihood of statements in a manner consistent with the tested model. In line with the experimental results, model fits and sensitivity analyses also provided very little evidence in favor of the model. This suggests that Bovens and Hartmann's present model fails to explain fully people's judgements in standard conjunction fallacy tasks.

*Keywords:* Conjunction fallacy; Source reliability; Bayesian models; Subjective probability

## 1. Introduction

The conjunction fallacy is one of the best-known judgment errors in the cognitive literature. The fallacy consists of judging the conjunction of two events as more likely than the least likely of the two events (Tversky & Kahneman, 1982). Thus, it appears that human judgment violates one of the most fundamental tenets of probability theory. Such a seemingly gross flaw in human rationality has prompted a wealth of empirical investigation. However, 26 years of extensive research[1] have failed to produce an adequate account of the phenomenon (Fisk, 2004).

Correspondence should be sent to Andreas Jarvstad, School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff, CF10 3AT, UK. E-mail: jarvstad@cardiff.ac.uk

Typically, the conjunction fallacy is elicited by asking participants to judge the probability that several statements describe a person. The best-known conjunction fallacy problem is the ''Linda'' problem: ''Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice and also participated in anti-nuclear demonstrations'' (Tversky & Kahneman, 1982, p. 92). The critical statements in the Linda scenario are ''Linda is a bank teller'' and the conjunction ''Linda is a bank teller and is active in the feminist movement.'' Approximately 85% rank the posterior probability of the conjunction in light of the description, P(bank teller, feminist|Linda), as more likely than the corresponding probability of the unlikely statement P(bank teller|Linda) (Tversky & Kahneman, 1982), thus exhibiting the conjunction fallacy.

Tversky and Kahneman (1982; see also Kahneman & Frederick, 2002) argued that the conjunction fallacy arises through the substitution of representativeness estimates for probability estimates. The representativeness of an evaluated object depends on ''... the degree to which it is (i) similar in essential properties to its parent population; and (ii) reflects the salient features of the process by which it is generated'' (Kahneman & Tversky, 1982, p. 33). ''Bank teller and feminist'' is more representative of the description of Linda than ''bank teller,'' and by attribute substitution the former is judged as more likely than the latter (Kahneman & Frederick, 2002). Most of the alternative accounts in the literature can be viewed as further elaborations of explanations also considered by Tversky and Kahneman (1982, 1983) such as estimating the wrong probabilities (Wolford, Taylor, & Beck, 1990), conversational implicatures (Dulany & Hilton, 1991), natural frequencies (Gigerenzer, 1996), potential surprise (Fisk & Pidgeon, 1998), and information value (Crupi, Fitelson, & Tentori, 2008; Massaro, 1994). However, neither representativeness, nor any of these alternative accounts, can arguably account for all moderating factors.

Several factors increase the incidence of the fallacy: designs where different participants rate each critical statement, asking participants to rank rather than rate statements, the insertion of foil statements between the critical statements, use of statements that are causally related (Tversky & Kahneman, 1983) and statements that are conditionally dependent (Fisk & Pidgeon, 1998).

Likewise, a range of factors decrease the incidence of the fallacy: the use of frequency formats, materials where the conjunction contains two unlikely components (Tversky and Kahneman, 1982, 1983), linguistic clarifications (Dulany & Hilton, 1991; Hertwig, Benz, & Krauss, 2008; Macdonald & Gilhooly, 1990; Wolford et al., 1990), the use of abstract problems, statistical sophistication on the part of the participants (Epstein, Denesraj, & Pacini, 1995), higher cognitive ability (Stanovich & West, 1998; but see Stanovich & West, 2008), and various training programmes (Agnoli & Krantz, 1989; Benassi & Knoth, 1993). However, only the removal of the description (of e.g., Linda) abolishes the fallacy (Tversky & Kahneman, 1983).[2]

Despite the existence of moderating factors, the conjunction fallacy is a very robust phenomenon. Tversky and Kahneman (1982, 1983) replicated it with a large number of participants and a variety of different manipulations and controls. In a recent review, it was concluded that ''an adequate account of the fallacy remains elusive'' (Fisk, 2004, p. 40) and so the search continues (e.g., Crupi et al., 2008; Hertwig et al., 2008; Wedell & Moro, 2008).

Given the apparent lack of an adequate account, Bovens and Hartmann's (2003) novel source reliability account is interesting. Source reliability seems important as it might be argued that most (if not all) sources of information are less than fully reliable. It would hence seem rational to take source reliability into account when reasoning with information provided by such sources. Intuitively, it seems appropriate that information be weighted by the reliability of the source (see also e.g., Hahn, Harris, & Corner, 2009; Harris & Hahn, 2009; Schum, 1981).

Bovens and Hartmann (2003) show that if a source of information is less than fully reliable there exist situations in which probability theory actually *prescribes* that the conjunction statement be judged as more likely than the least likely component statement. Specifically, receiving a report that matches our prior belief (a report of the likely fact $REP_L$), from a source whose reliability we are agnostic about, can cause greater belief updating than receiving a report that seems improbable given our prior belief (a report of the unlikely fact $REP_U$). This fact can cause the belief in the conjunctive statement ($REP_{LU}$) to be higher than the belief in the unlikely statement presented on its own ($REP_U$)—rendering the fallacy a ''non-fallacy'' (see ''Bovens and Hartmann's Source Reliability Account'' below for a more detailed exposition; see also McKenzie, Wixted, & Noelle, 2004 and Corner, Harris, & Hahn, 2010, for other examples of explaining away apparent irrationality using source reliability).

The idea that the conjunction fallacy stems from an otherwise rational process is not new: It has been argued that the fallacy is due to participants following conversational norms (e.g., Adler, 1984), that the term probability cues nonmathematical reasoning (e.g., Hertwig & Gigerenzer, 1999), and that single components are interpreted to imply the negation of other single components (e.g., Dulany & Hilton, 1991). It may, for example, not be fallacious to rate P(bank teller, feminist|Linda) higher than P(bank teller, ¬feminist|Linda).

However, these claims themselves have been challenged on both theoretical and empirical grounds (see e.g., Donovan & Epstein, 1997; Sides, Osherson, Bonini, & Viale, 2002; Tentori, Bonini, & Osherson, 2004). Thus, while some errors of judgment are likely to be due to participants interpreting conjunction problems differently than intended by experimenters, this is unlikely to be the only cause.

Bovens and Hartmann's (2003) account is interesting because it too explains the controversial nonnormative results by a normative process. However, it differs from previous rational process accounts (cf., e.g., Crupi et al., 2008; Massaro, 1994) in that it does not assume that participants evaluate the *wrong* property given the experimental instructions (e.g., confirmation, Crupi et al., 2008), only that participants take into account the reliability of sources. In this regard, it is an appealingly parsimonious account. Moreover, it assumes not only a rational process underlying the conjunction response; it argues that this response itself, properly considered, can actually be correct. Hence, it merits closer theoretical and empirical scrutiny.

As we will outline below, Bovens and Hartmann (2003) demonstrate that the fallacy *can* be a normative response. In establishing this result, Bovens and Hartmann make use of a specific model. This model may simply be viewed as an analytical tool with which an existence proof of this interesting and important normative fact about probability judgments is obtained. However, one can also take this model seriously as a potential account of what people actually do when faced with standard conjunction fallacy problems. In other words, this instantiation of the source reliability account can also be evaluated for empirical adequacy.

In order to assess the descriptive validity of the model, three experiments were carried out. In two of these experiments, model parameters were manipulated and the effect on the incidence of the conjunction fallacy was assessed. The remaining experiment tested people's ability to infer source reliability from the prior probability of statements, a critical assumption in Bovens and Hartmann's account. In addition to these empirical tests, the ability of the model to fit participant data was evaluated. Finally, Bovens and Hartmann's source reliability model was subjected to a sensitivity analysis.

To anticipate our results, it appears that people are able to infer source reliability from the prior probability of statements as the model predicts. A person who states that ''Linda is a bank teller'' produces lower trust than a person who states, ''Linda is a bank teller and a feminist.'' However, the empirical tests and the modeling conducted suggest that Bovens and Hartmann's present model struggles to explain fully people's tendency to rate ''bank teller and feminist'' as more probable than ''bank teller.''

## 2. Bovens and Hartmann's source reliability account

Bovens and Hartmann's (2003) formulate a highly generic and noncommittal source reliability model, which they apply to a broad range of philosophical topics in epistemology, and the philosophy of science. Their model embodies the idea that the partially reliable sources that we might encounter in everyday life report accurately on the state of the world when reliable, and when unreliable, deviate in, what are to us, effectively random ways because nothing more specific is known.

The model is illustrated by the Bayesian Network (BN) with binary nodes in Fig. 1A). X stands for a claim that is either true or false and REL indicates whether the source is reliable

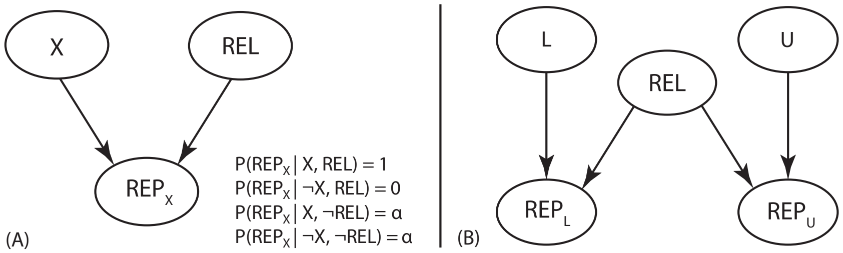
X
REL
$REP_X$
$P(REP_X \mid X, REL) = 1$
$P(REP_X \mid \neg X, REL) = 0$
$P(REP_X \mid X, \neg REL) = \alpha$
$P(REP_X \mid \neg X, \neg REL) = \alpha$
(A)
L
U
REL
$REP_L$
$REP_U$
(B)

Fig. 1. (A) Bovens and Hartmann's generic source reliability model as a Bayesian Network, with a partial conditional probability table. (B) Bovens and Hartmann's model of the conjunction fallacy (both adapted from Bovens and Hartmann, 2003).

or not. One's prior degree of belief in X and REL is represented by x and $\rho$, respectively. These prior beliefs determine jointly the influence that the report that X obtains ($REP_X$, or that X does not obtain $\neg REP_X$[3]) will have on our posterior belief in X and REL. As illustrated by the conditional probabilities in Fig. 1A), a reliable source (REL) is believed to report X if and only if X is true. This means that if the source is reliable it is perceived as telling the truth. On the other hand, if the source is unreliable (¬REL), the source is believed to produce a report X with probability $\alpha$, where a value of .5 reflects ignorance about how and why an unreliable source generates its reports.

Fig. 1B shows Bovens and Hartmann's application of this generic model to the conjunction fallacy problem. X has been replaced with the nodes L and U. These represent a likely and an unlikely state of the world, respectively (e.g., ''Linda is bank teller'' and ''Linda is a feminist''). $REP_L$ and $REP_U$ are report nodes and REL is the reliability node, with the same interpretation as in Fig. 1A.[4] The personality description (e.g., the description that makes Linda sound like a person who is likely to be a feminist, but unlikely to be a bank teller) is assumed to set the prior probabilities of L and U.

Here, as in Bovens and Hartmann (2003) conceptualization of the conjunction fallacy, it is assumed that $\alpha = .5$, because nothing further is known about the potential source. However, the critical implication that rating LU as more likely than U is sometimes normative is not specific to parameter values of .5. Moreover, the impact of changing the values of this parameter will be explored in the section ''Sensitivity analyses'' below.

Intuitively, the model works as follows. Assume that we are agnostic about the reliability of a source ($\rho = .5$), and that we are told ''Linda is a bank teller'' (which we consider unlikely given the description). Then, the provided report will not only cause a relatively small change in our belief that ''Linda is a bank teller'' (because we do not consider the source to be particularly reliable), it will also decrease our belief in the reliability of the source (because the statement does not match our prior belief). The interesting situation occurs when a source provides both a likely and an unlikely report. Then, the additional likely report will cause both an increase in the posterior belief and an increase in the perceived reliability of the source (relative to an unlikely report on its own). Thus, source reliability interacts with prior beliefs and reports, creating situations in which the probability that two reports are true can be greater than the probability than the unlikeliest report on its own is true.

In the model (Fig. 1B), the prior beliefs (in e.g., L) is independent of other prior beliefs (e.g., U), *conditional* on the personality description. This does not imply that bank teller and feminist are otherwise independent. Being a bank teller may well decrease the probability of being a feminist. Instead, it means that given what we know about Linda (namely that she is the kind of person who is unlikely to be a bank teller and likely to be a feminist) receiving reports about one of these specific beliefs does not further make the other belief more or less likely.[5]

Exploiting the conditional dependencies in Fig. 1B, Bovens and Hartmann (2003) derive the following equation for the posterior probability of the unlikely component statement:

$$P(U|REP_U) = \frac{\text{priorU}(\rho + \alpha\neg\rho)}{\text{priorU}\rho + \alpha\neg\rho} \tag{1}$$

and for the posterior probability of the conjunction:

$$P(L, U|REP_L, REP_U) = \frac{\text{priorLpriorU}(\rho + \alpha^2\neg\rho)}{\text{priorLpriorU}\rho + \alpha^2\neg\rho} \tag{2}$$

They then show that, given a prior reliability of .5, there is a considerable range of priors for which the posterior degree of belief in the conjunction will be greater than for the unlikely component. That is, the difference ΔP between the posterior of the conjunction and that of the unlikely statement alone will be greater than zero:

$$\Delta P = P(L,U|REP_L, REP_U) - P(U|REP_U) > 0 \tag{3}$$

Normatively, wherever ΔP is positive, people *should* rate the conjunction as more likely than the unlikely component, and the ‘‘conjunction fallacy’’ is not a fallacy. If ΔP is negative, rating the conjunction as more likely than the least likely statement remains fallacious.

Fig. 2 depicts the ΔP ‘‘landscape’’ for a range of priors (priorL ≥ .5001, priorU ≤ .4999) for $\rho$ and $\alpha = .5$. The range of priors over which it is normative to commit the ‘‘fallacy’’ is shown by the area of positive ΔP space (grey area). As can be seen, the seeming fallacy is the normative response over a wide range of priors.

For the purpose of empirical model tests, qualitative model predictions of fallacy rates (the proportion of people committing the fallacy) can be derived by relating the area of positive ΔP space to fallacy rates. Given variation in people’s prior beliefs in the statements (i.e., prior beliefs are distributed over the space in Fig. 2), and everything else being equal, one would expect a greater area to produce greater fallacy rates (simply because an increase in area increases the likelihood of a given data point being in that area).

Are people influenced by source reliability and prior probabilities when judging the likelihood of statements? Empirical studies on the effects of source reliability are carried out

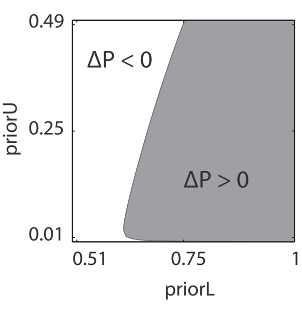

Fig. 2. Difference landscape (ΔP) over priorU = .0001–.4999, priorL = .5001–.9999 for $\rho = .5$ and $\alpha = .5$. The white area depicts priors over which the conjunction fallacy is indeed fallacious (−ΔP). The grey area depicts priors over which the ‘‘fallacy’’ is the normative response (+ΔP).

mainly in persuasion research (e.g., Eagly & Chaiken, 1993; Pornpitakpan, 2004). In this context, source credibility rather than source reliability is studied. Typically, credibility is viewed as a two-factor construct, with source expertise and source trustworthiness as factors (Pornpitakpan, 2004; but see Birnbaum & Stegner, 1979, who argue for a three-factor construct).

The dominant framework of dual route models of persuasion (for an overview see Eagly & Chaiken, 1993; but see Kruglanski & Thompson, 1999) has meant that persuasion research has typically (though not exclusively, see e.g., Chaiken & Maheswaran, 1994) focussed on recipients processing either message content *or* source characteristics (if motivated they will process the former, if unmotivated the latter). By contrast, Bovens and Hartmann's (2003) model prescribes detailed interactions between message content and source characteristics (see also Hahn et al., 2009 for an empirical verification of such interactions).

In this sense, Bovens and Hartmann's (2003) source reliability model is more sophisticated than source credibility accounts. In other aspects, however, their model embodies a much simpler notion of source reliability. In many real-world contexts, one might have far more detailed insight into how and why a particular source could potentially deviate from the truth. Nevertheless, the model's conceptualization suffices to make the general point: The conjunction fallacy may be normative under certain conditions. Other aspects of source reliability (such as source expertise) could readily be incorporated within its general probabilistic framework.

## 3. Experiment 1

The source reliability account predicts that the addition of extra components will affect the incidence of the conjunction fallacy (Bovens & Hartmann, 2003, p. 88). Fig. 3 shows the model with an added third component (A and $REP_A$). Intuitively, an additional unlikely (likely) component decreases (increases) the perceived reliability of the source and decreases (increases) the belief in the conjunction vis-à-vis conjunctions without the extra statement.

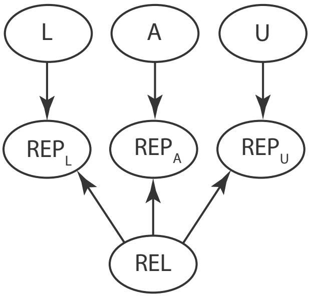

Fig. 3. Bovens and Hartmann's model for a three-component conjunction problem.

Given the conditional independencies, it is straightforward to extend Bovens and Hartmann's equations to accommodate the extra component:

$$P(L, A, U|REP_L, REP_A, REP_U) = \frac{\text{priorLpriorApriorU}\left(\rho + \alpha^3 \neg\rho\right)}{\text{priorLpriorApriorU}\rho + \alpha^3 \neg\rho} \tag{4}$$

For the three-component problem, the normativity of answers is likewise assessed by the difference function, now based on the three-component conjunction and the least likely component P(U|REP$_U$):

$$\Delta P = P(L, A, U|REP_L, REP_A, REP_U) - P\left(U|REP_U\right) > 0 \tag{5}$$

As above (Eq. 3), a positive ΔP value means that the ''fallacy'' is the normative response.

Fig. 4 shows the result of comparing a three-component network (Fig. 3) to the original two-component network (Fig. 1B). Specifically, it shows the normalized area of parameter space in which the fallacy is the normative response (+ΔP$^2$), for each network, as a function of the prior probability of the additional statement (''A'' in Fig. 2). As outlined above, this area can be related to the rate of the fallacy.

For a source reliability of .5 (Panel 2, Fig. 4), the model predicts that three component problems with an added likely component should result in a greater number of participants committing the fallacy—compared to standard two component problems (for priorA >∼ .6, LLU > LU). Conversely, three component problems with added unlikely components result in a smaller area relative to standard problems and should therefore result in fewer fallacies (LUU < LU). Perhaps the strongest prediction is that three component problems with additional unlikely components will produce a lower fallacy rate than three component problems with additional likely components (LUU < LLU). Fig. 4 also illustrates these relationships for less- (Panel 1) and more-reliable sources (Panel 3). As can be seen, the relationship between the two- and three-component areas change slightly as a function of source

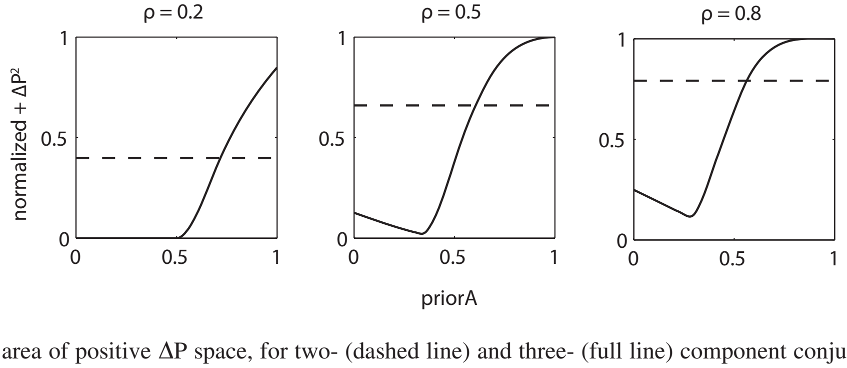

Fig. 4. The area of positive ΔP space, for two- (dashed line) and three- (full line) component conjunction problems, as a function of the prior probability of the additional component (priorA), for increasing source reliability parameter values (panels 1–3). As there is no added component in two-component problems the area is constant. Fig. 4 is defined over priorU in the range .0001–.4999 and for priorL in the range .5001–.9999 and α = .5.

reliability. Note, however, that the strong prediction that LUU < LLU remains across changes in source reliability.

To our knowledge, only three studies with three-component conjunctions have been conducted. Teigen, Martinussen, and Lund (1996a, experiment 1) found no evidence that the addition of an extra likely component (LLU) increased the incidence of the fallacy relative to a two-component conjunction (LU) for a classical conjunction fallacy problem (''Linda,'' Tversky & Kahneman, 1983). In a second experiment, various outcomes in the 1994 football World Cup were estimated. When evaluating these outcomes, three-component conjunctions resulted in fewer conjunction fallacies compared to two-component conjunctions (see also Teigen, Martinussen, & Lund, 1996b). Unfortunately, these results do not speak to the predictions of Bovens and Hartmann's (2003) model. The frequency of two-component fallacies in the above studies was an aggregate of the incidence of the fallacy for three two-component conjunctions (e.g., $L_1U$, $L_1L_2$, $L_2U$). Hence, it is impossible to determine which two-component conjunction the three-component conjunction fallacy (e.g., $L_1L_2U$) frequency changed relative to.

A third study was conducted by Stolarz-Fantino, Fantino, Zizzo, and Wen (2003). They used a standard conjunction problem (''Bill,'' Tversky & Kahneman, 1983) but participants were required to estimate only the conjunction and were given the component probabilities ($P(L) = .8$, $P(U_1) = .2$, $P(U_2) = .1$). These authors failed to find an effect of an additional likely component on the rate of the fallacy (LUU = 55% incidence, LU = 52% incidence). However, it is unclear whether explicit probabilities are processed in the same way as internally generated probabilities. In summary, it is difficult assess the predictions of Bovens and Hartmann's (2003) model on the basis of previous results.

To investigate if an extra component does indeed affect the incidence of the conjunction fallacy, the probability of an extra component in a conjunction fallacy problem (''Helen,'' Fisk & Pidgeon, 1996) was manipulated. After reading a personality description, participants' either rated an LLU, an LU, or an LUU conjunction and their respective component probabilities. If Bovens and Hartmann's (2003) model explains the fallacy, then one would expect fallacy rates to be lower in LUU than in LU problems and rates to be lower in LUU problems than in LLU problems. One might also expect fallacy rates to be higher for LLU than LU problems.

### *3.1. Method*

#### *3.1.1. Participants*

Eighty-nine undergraduates at Cardiff University participated for course credit.

#### *3.1.2. Materials*

The material was presented in a questionnaire leaflet. Participants whose answers were incomplete or participants who reported having knowledge about the conjunction fallacy were excluded ($N = 14$).

There were three versions of a modified Helen problem (Fisk & Pidgeon, 1996) corresponding to three conditions. The LU version contained one likely (L, ''Helen loves going to parties'') and one unlikely (U, ''Helen collects stamps for a hobby'') component statement and their conjunction (LU). The LLU and the LUU version contained an additional likely (L, ''Helen is a holiday rep.''), and an additional unlikely (U, ''Helen is a post office clerk'') component, respectively. The LLU version is provided as an example:

> Helen is a sociable person who thrives on human company and seeks out exciting activities. She is restless, impulsive, and optimistic.
>
> Rate the chance that each of the following statements applies to Helen. Use a scale from 0 to 100 to make your rating. 0 on the scale indicates ''definitely no,'' 100 indicates ''definitely yes,'' and 50 indicates a ''50/50 chance.'' You are free to use any whole number between 0 and 100.
>
> Helen loves going to parties _______
>
> Helen is a holiday rep. ______
>
> Helen collects stamps for a hobby _______
>
> Helen loves going to parties, is a holiday rep, and collects stamps for a hobby ______

### *3.1.3. Design and procedure*

A mixed design was used. Conjunction type (LU, LLU, or LUU) was manipulated between subjects and statement type (L, U, LU, etc.) was manipulated within subjects. The dependent variable was the chance estimate. Participants completed the questionnaires individually in small groups. Of the 75 participants, 22 were assigned to LU, 28 to LLU, and 25 to LUU.

## *3.2. Results*

As is evident from Fig. 5, the manipulation of the perceived likelihood of the statements (transformed to standard probabilities) describing Helen appeared successful. Likely statements appear to have been rated as likely and unlikely statements as unlikely. A relatively high average rating for the conjunction statements is suggestive of a general conjunction fallacy effect.

To assess whether likely statements were rated as likely, and unlikely statements were rated as unlikely, one-sample *t* tests comparing the mean rating of single statements to the reference P(Statement) = .5 were conducted. As can be seen in Table 1, all likely statements were, on average, rated as likely (>.5), and all unlikely statements were, on average, rated as unlikely (<.5).

A nominal conjunction fallacy incidence variable was created: Participants who conformed to (LU/LLU/LUU) > (U) were classified as exhibiting the fallacy and participants who conformed to (LU/LLU/LUU) ≤ (U) as not exhibiting the fallacy (where U is the least likely component statement for each participant).

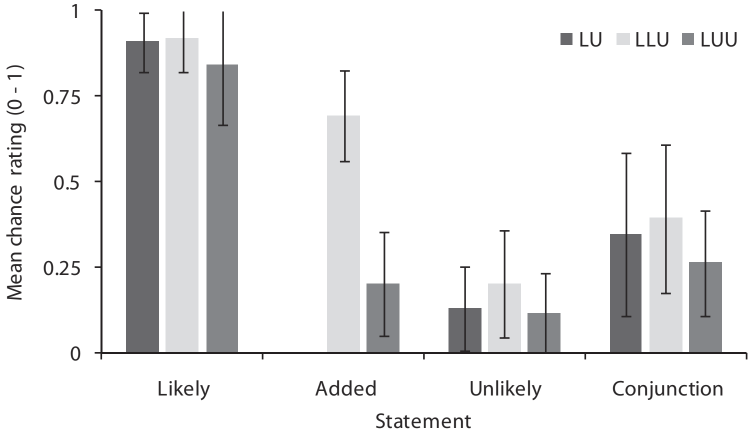

Fig. 5. Mean chance rating as a function of condition and statement type (error bars are ±1 standard deviation). Note that for the ''Added Statement'' the light grey bar indicates a likely statement and the dark grey bar indicates an unlikely statement.

Table 2 shows that the rate at which fallacies were committed did not differ much across conditions. The maximal difference was between those who received the standard problem (LU) and those who received an extra unlikely statement (LUU) and this apparent difference was in the opposite direction to the one predicted.

A chi-square contingency analysis found no evidence that the frequency of the fallacy depended on condition [$\chi^2(2, N = 75) = .362$, $p = .834$]. This analysis is supported by a comparison of fallacy rates across conditions by means of Bayes Factors (Kass & Raftery, 1995; Lee & Wagenmakers, 2005). While there are many principled arguments for Bayesian statistics, the most important one for the present purposes is that Bayesian methods allow inferences in favor of the null hypothesis (Gallistel, 2009; Rouder, Speckman, Sun, Morey,

Table 1
Comparisons of the mean ratings for single statements to an equally likely/unlikely reference (P [Statement] = .5)

| Condition | Statement | *t*-value | *df* | *p* | Cohen's *d* |
|---|---|---|---|---|---|
| LU | L | 22.15 | 21 | 5.E-16 | 4.73 |
| | U | −14.17 | 21 | 3.E-12 | 3.02 |
| LLU | L | 21.88 | 27 | 1.E-18 | 4.14 |
| | Added L | 7.71 | 27 | 3.E-08 | 1.46 |
| | U | −10.17 | 27 | 1.E-10 | 1.92 |
| LUU | L | 9.58 | 24 | 1.E-09 | 2.34 |
| | Added U | −9.92 | 24 | 6.E-10 | 1.30 |
| | U | −16.67 | 24 | 1.E-14 | 2.59 |

*Note:* All effects remain significant when Bonferroni corrected.

Table 2
Frequency of fallacies and inferences about differences in rates across conditions

| Frequency of Fallacies | | | | Fallacy Rates Comparisons | | |
|---|---|---|---|---|---|---|
| | LU | LLU | LUU | | %difference | Bayes Factor |
| % of fallacies | 68.2 | 71.4 | 76.0 | LU vs. LLU | 3.3 | 3.1 |
| No. of fallacies | 15 | 20 | 19 | LU vs. LUU | 7.8 | 2.7 |
| Total *N* | 22 | 28 | 25 | LLU vs. LUU | 4.6 | 3.2 |

*Note:* Bayes Factors express the odds in favor of null hypothesis of no difference between fallacy rates.

& Iverson, 2009). Bayes Factors can straightforwardly be interpreted as odds in favor of one hypothesis over another. Here, we adopt the recommendation by Jeffreys (1961) and view odds greater than 3 as ''some evidence,'' odds greater than 10 as ''strong evidence,'' and odds greater than 30 as ''very strong evidence.''

The computed Bayes Factors suggest that the null hypothesis is between 2.7 and 3.2 times more likely than the hypothesis that the rates differ. Thus, some evidence in favor of the null hypothesis for the LU and LLU comparison, and for the LLU and the LUU comparison, was found. The evidence for the LU and LUU comparison, on the other hand, was inconclusive.

### *3.3. Discussion*

To reiterate, conjunction fallacies were generally expected to be more numerous, relative to a standard conjunction scenario (LU) when a likely component was added (LLU) and less numerous when an unlikely component was added (LUU). Importantly, fallacies were expected to be less numerous in the LUU condition than in the LLU condition. Experiment 1 did not confirm these predictions. There was some evidence that the fallacy rate did not differ for two of the comparisons (LU vs. LLU and LLU vs. LUU) and there was insufficient evidence to favor either the hypothesis of a difference, or the hypothesis of sameness, for one of the comparisons (LU vs. LUU).

An earlier version of this experiment,[6] with a different conjunction problem ($N = 60$), likewise found no evidence that fallacy rates differed as a function of added likely or unlikely statements. When that dataset is combined with the dataset of Experiment 1, the resulting Bayes Factors provide some support for the null hypothesis for all three comparisons (Bayes Factors: 4, 3.9, and 3.7). This, in conjunction with the previous null-finding of Stolarz-Fantino et al. (2003), makes it unlikely that the addition of a third moderately likely or moderately unlikely component to standard conjunction problems affects the conjunction fallacy.

Note that the failure to find an effect of added components is problematic for other accounts also. For example, it seems reasonable to assume that LLU conjunctions are more representative of the personality description than LUU conjunctions are. According to representativeness the former should thus result in more fallacies than the latter and this did not occur.

From the perspective of Bovens and Hartmann's (2003) model, one possible explanation for the lack of an effect on the incidence of the fallacy is that people are incapable of extracting source reliability from conjunction scenario statements. In other words, Bovens and Hartmann's (2003) model may have failed because people cannot (or will not) infer source reliability from the prior likelihood of statements. This possibility was explored in Experiment 2.

## 4. Experiment 2

In order to test the possibility that people cannot (or do not) extract source reliability estimates from the likelihood of statements, participants were asked to rate the reliability of fictional sources that provide standard conjunction fallacy scenario statements. If people are able to infer source reliability from the probability of statements, one would expect source reliability ratings to be ordered as follows: L > LU > U (Bovens & Hartmann, 2003). That is, statements that are likely, given known information, should cause high source reliability ratings. Unlikely statements should cause low reliability ratings. For conjunctions, the high probability statement should increase reliability ratings and the low probability statement should decrease reliability ratings, causing the overall reliability rating to be between the two statements made separately.

### *4.1. Method*

#### *4.1.1. Participants*

Ninety psychology undergraduates at Cardiff University participated in exchange for course credit.

#### *4.1.2. Materials*

Three questionnaires were used. Each questionnaire contained a personality description of the fictitious character Bill. Each questionnaire also described someone making a statement about Bill's character. The provided statement differed across the three questionnaires. It was either ''Bill plays jazz for hobby,'' ''Bill is an accountant,'' or ''Bill is an accountant and plays jazz for a hobby.'' Participants were asked to indicate how reliable the person making these statements seemed. Reliability was rated on a scale between 0 (not at all reliable) and 100 (completely reliable). The unlikely component scenario follows as an example:

> You know that Bill is 34 years old. He is intelligent, but unimaginative, compulsive and generally lifeless. In college, he was strong in mathematics but weak in social studies and literature.

Now somebody tells you that Bill plays jazz for a hobby.

How reliable would you judge this person to be?

Please provide an estimate between 0 (not at all reliable) and 100 (completely reliable).

### *4.1.3. Design and procedure*

A between-subjects design was used. Participants were randomly allocated to one of the three conditions. Thirty participants rated the likely component statement, 31 participants rated the unlikely component statement, and 29 participants rated the conjunction.

### *4.1.4. Data analysis*

Because of outliers, the data were trimmed by ∼14% ($N = 4$) in each condition. The presence of extreme outliers (e.g., .09 in the likely condition) was presumably due to the between-subject design. That is, without other statements acting as comparison statements (anchors) a greater variation in the data is likely. A Welch's variance-weighted ANOVA was carried out in conjunction with Tamhane post-hoc tests.

## *4.2. Results*

As can be seen in Fig. 6, the general pattern of reliability ratings was as predicted. Source reliability had an overall effect on reliability ratings [F(2, 48.9) = 16.41, $p < .001$, MSE = .021]. Furthermore, the source was rated as more reliable when providing the likely component statement (''Accountant'') than when providing the unlikely statement (''Jazz player'') (mean difference = .24, $p < .001$, $SE = .04$). The source was also rated as more reliable when providing the likely component statement than when providing the conjunction (''Accountant & Jazz player'') (mean difference = .12, $p = .002$, $SE = .03$). Finally, the source providing the unlikely statement (''Jazz player'') was rated as less reliable than

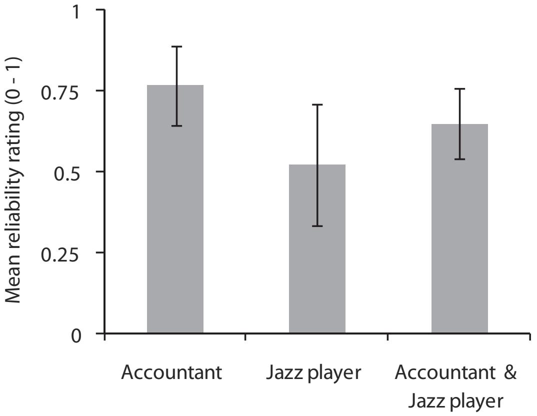

Fig. 6. Mean reliability ratings as a function of statement type. Error bars are ±1 standard deviation.

the source providing the conjunction statement (mean difference = −.13, $p$ = .014, $SE$ = .04).

### 4.3. Discussion

The prior likelihood of statements affected source reliability judgments. Thus, a person who provides a likely statement (L) is judged to be more reliable than someone who provides an unlikely statement (U). A person who provides a statement composed of a likely and an unlikely statement (LU) is perceived to be more reliable than a person who provides unlikely statements (U), but less reliable than a person who provides likely statements (L).

It is interesting that source reliability ratings in Experiment 2 were similar to the probability estimates in Experiment 1 (see LU condition, Fig. 5 above, see also e.g., Fisk & Pidgeon, 1996). A high similarity between ratings of apparently different properties has been taken as evidence that one property is being used as a substitute for the other. For example, when drafting proposals for experimental tests of their confirmation account, Crupi et al. (2008) suggest that a match between confirmation ratings and probability ratings would provide support for their model. Similarly, Fisk (2002) has argued that the strong correlation between measures of surprise and measures of probability is supportive of a surprise account of the fallacy. Kahneman and Frederick, citing a high correlation between mean rankings of similarity and mean rankings of probability, state: ''The correlation between representativeness and probability is nearly perfect (.97). *No stronger* support for attribute-substitution could be imagined'' (Kahneman & Frederick, 2002, p. 61, emphasis added).

It appears that the above authors believe that a high correlation between *their* chosen measure and probability ratings indicates that people use *their* measure (and not probability, or any other measure). Although we cannot perform a reliable correlation analysis on our data, we conjecture that a high positive correlation would be found between source reliability ratings and probability ratings.

The very multitude of correlations that have been proposed (Crupi et al., 2008), or found (Fisk, 2002; Tversky & Kahneman, 1983), however, suggests that the identification logic might be flawed. A sceptic could argue that whatever it is that participants *do* evaluate is simply more or less independent of the specific measure that they are being asked to use. At the same time, a more positive interpretation of such correlations exists: If a single underlying psychological property (e.g., subjective probability) is used to derive the proposed alternative measures, a similar result may obtain. Either way, caution in interpreting correspondences between measures or properties seems appropriate.

In conclusion, Bovens and Hartmann's (2003) intuition that the prior probability of statements influences source reliability seems correct. Hence, the lack of an effect in Experiment 1 is unlikely to have been due to an inability to extract source reliability from the prior probability of statements. While the absence of an effect would have ruled out Bovens and Hartmann's account, the positive observation of the predicted correspondence, though encouraging, provides, in and of itself, only rather weak support for the account.

Despite the current results, the account may be partially true if source reliability affects the frequency of the conjunction fallacy but is not the main determinant of it. In an attempt to produce a more sensitive test, we explicitly manipulated source reliability in Experiment 3.

## 5. Experiment 3

If, as Bovens and Hartmann (2003) argue, the conjunction fallacy arises due to the normative combination of subjective probabilities and perceived source reliability, different conjunction fallacy rates are expected for reliable versus unreliable sources. As can be seen in Fig. 7, the range of prior beliefs over which the fallacy is normative increases with increasing reliability (except for the degenerate case when sources are perfectly reliable). Given reasonable variability in people's prior beliefs, a reliable source should thus lead to an increase in the frequency of the fallacy relative to an unreliable source.

Both Bovens and Hartmann's (2003) model and source credibility studies (Pornpitakpan, 2004) lead to the additional prediction that all statements will be rated as more probable when originating from reliable sources than from unreliable sources. Thus, a second prediction was that source reliability would moderate believability estimates. In Experiment 3, these predictions were tested by manipulating the reliability of the source of standard conjunction fallacy statements.

### *5.1. Method*

#### *5.1.1. Participants*

Eighty-two undergraduates at the School of Psychology, Cardiff University, participated for course credit.

#### *5.1.2. Materials*

The material was presented in questionnaire format. Each questionnaire contained one modified Bill scenario (Tversky & Kahneman, 1983) and one modified Helen scenario (Fisk & Pidgeon, 1996). Both scenarios contained a short personality description

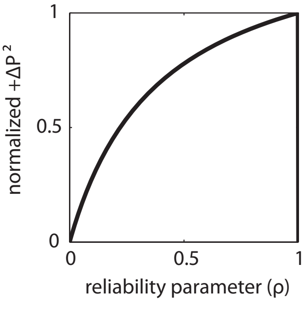

Fig. 7. Normalized +ΔP area as a function of source reliability for priors L and U in the range .5001–1 and .0001–.4999, respectively, and α = .5.

and a statement that the participant speaks to James (for Bill) or Mary (for Helen) who is either quite reliable or quite unreliable.[7] The scenarios contained two-component statements, one likely (L) and one unlikely (U), and their conjunct (LU). Finally, each scenario contained the instruction to provide an estimate between 0 and 100 as to how believable each statement is. The personality description and the statements of the Helen problem were identical to that of the LU condition in Experiment 1. The Bill problem is provided below:

> Bill is 34 years old. He is intelligent, but unimaginative, compulsive, and generally lifeless. In university, he was strong in mathematics but weak in social studies and literature.
>
> You speak to James who knows Bill well and is quite reliable. Reliable James tells you one of the following:
>
> Bill is an accountant. _____________
>
> Bill plays Jazz for a hobby. ______________
>
> Bill is an accountant and plays Jazz for a hobby. _____________
>
> How much would you believe the statement in each case? Take into consideration what you know about Bill and the reliability of James. Please provide a rating between 0 (definitely untrue) and 100 (definitely true) for each statement.

### *5.1.3. Design*

A mixed Latin square design was used. Source reliability was manipulated between participants (reliable, unreliable) and statement type (L, U, LU) was manipulated within participants. Each participant rated one of the scenarios with statements from a reliable source, and the other scenario with statements from an unreliable source. The order of the scenario/reliability combination and the order of component statements were counterbalanced across participants. Knowledge of the conjunction fallacy was controlled for. Participants were queried about whether they had heard about the Linda problem or the conjunction fallacy before. If they responded affirmatively, they were asked if they saw it as applicable to the present questionnaire. If responding affirmatively a second time, they were asked to give a short description of the problem/fallacy.

### *5.1.4. Procedure*

Participants rated the two scenarios individually in small groups. The task took approximately 5 min.

### *5.1.5. Data analysis*

All responses were transformed to standard probabilities. A nominal conjunction fallacy variable was created by assigning estimates that conformed to $P(L, U) > P(U)$ as fallacious and estimates that conformed to $P(L, U) \leq P(U)$ as nonfallacious.

Two participants were excluded from the analysis because they did not exhibit the conjunction fallacy, stated that they knew about the fallacy, saw it as applicable to the problem, and could describe it fairly well. Five other participants responded affirmatively to the control question but could not describe it accurately. These participants were not excluded from the analysis.

There were no order effects of the component statements or the order of source reliability presentation. Data were collapsed across these counterbalancing measures. The procedure resulted in 40 P(L), P(U), and P(L, U) estimates for reliable statements and 40 P(L), P(U), and P(L, U) estimates for unreliable statements, for each of the two scenarios.

One mixed ANOVA, with source reliability as a between-subjects factor and statement type (L, U, LU) as a within-subject factor, was computed per scenario. Where needed, Greenhouse-Geisser corrections were applied. One chi-square test (per scenario) was computed to assess whether the frequency of the fallacy differed as a function of source reliability. Additionally, Bayes Factors were computed for differences in rates of the fallacy as a function of source reliability.

### 5.2. Results

As can be seen in Fig. 8, reliable sources produced significantly higher believability ratings than unreliable sources for both the Bill [F(1, 78) = 15.46, $p < .001$, $\eta_p^2 = .17$] and the Helen scenario [F(1, 78) = 27.89, $p < .001$, $\eta_p^2 = .26$]. Thus, the source reliability manipulation was effective. Furthermore, there was an overall effect of statement type on believability ratings for both scenarios [Bill, F(1.74, 135.36) = 115.81, $p < .001$, $\eta_p^2 = .59$; Helen, F(1.76, 137.07) = 224.39, $p < .001$, $\eta_p^2 = .74$]. There was also a source reliability

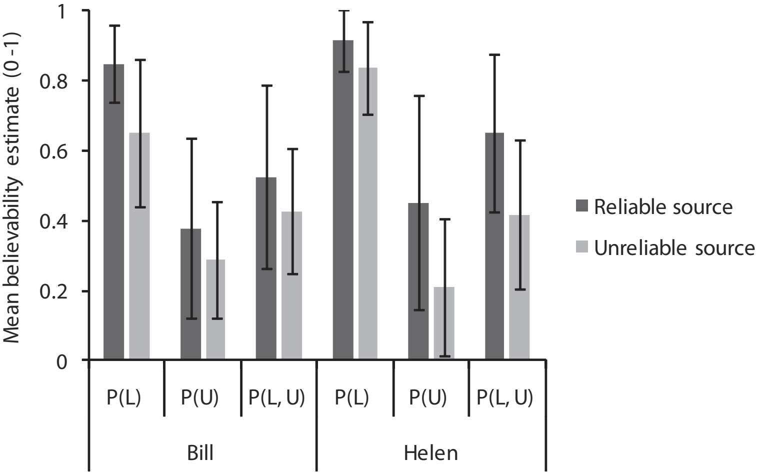

Fig. 8. Mean believability estimate as a function of scenario type, statement type, and source reliability. Error bars are ±1 standard deviations.

and statement type interaction for the Helen scenario [F(1.76, 137.07) = 6.22, $p$ = .004, $\eta_p^2$ = .074].

For the Bill scenario, 70% (28) committed the fallacy when rating statements from a reliable source, compared to 67.50% (27) when rating statements from an unreliable source [$\chi^2$(1, $N$ = 80) = .058, $p$ = .81]. For the Helen scenario, 72.5% (29) committed the fallacy whether rating statements from reliable or unreliable sources [$\chi^2$(1, $N$ = 80) = .000, $p$ = 1]. Thus, fallacy rates do not appear to change as a function of source reliability. Computed Bayes Factors indicate some evidence in favor of the hypothesis of no difference in fallacy rates between sources of low and high reliability (Bill = 3.8, Helen = 4.1).

### *5.3. Discussion*

In Experiment 3, source reliability was explicitly manipulated by asking participants to rate standard conjunction fallacy scenarios as if the statements were coming from reliable or unreliable sources. As predicted, source reliability affected overall ratings. Reliable sources produced higher ratings compared to unreliable sources. However, the analysis of the conjunction fallacy rates did not produce significant results. Instead, some evidence for the null hypothesis was found: Conjunction fallacy rates do not appear to be influenced by source reliability. The failure to find an effect is unlikely to be due to a weak experimental manipulation given that source reliability affected overall ratings.

Although other accounts do not generally factor in source reliability, reasonable predictions can be made. Changing the reliability of the source arguably does not change the representativeness (Tversky & Kahneman, 1983) of statements. In this regard, representativeness correctly predicted the absence of an effect on the frequency of the conjunction fallacy. However, the fact that reliable sources result in higher overall ratings, relative to unreliable sources, suggests that participants are not simply substituting representativeness for probabilities (cf., Kahneman & Frederick, 2002). If ratings were based on representativeness alone, it should arguably not matter if an unreliable person reports that Linda is a ''feminist bank teller'' or whether a reliable person reports the same.

## 6. Modeling the source reliability account

Experiment 1 and Experiment 3 provided some evidence against Bovens and Hartmann's (2003) model. It is important to note, however, that the predictions that these experiments were based on were in turn based on assumptions about participants' belief in the reliability of the source. It is possible that the failure to find support for the model was due to participants' beliefs being located in regions of parameter space where the predictions do not hold and/or where the power to detect effects was very low.

In the following, we take into account participants' belief in the reliability of the source, as it is conceptualized in the model, by fitting the source reliability model to participants' data directly. If the predictions for the experiments were simply mismatched to participants' beliefs, then the model should provide excellent fits when it is allowed to choose the best fitting reliability parameter. In order to make the quantitative fit more meaningful, we also compare the model to two averaging models.

## 6.1. Modeling methods

### 6.1.1. Comparison models

To our knowledge, there are only a few conjunction fallacy accounts that are able to make quantitative predictions given ratings of statements L and U. Wyer's (1976) model is one of the models that can be fit given only L and U data. Wyer developed the model to account for conjunctive probability estimates, which he found to be poorly fit by a normative multiplicative model as conjunctive probabilities are overestimated (Wyer, 1970). Wyer's model calculates the mean of a probability average and a probability product:

$$\mathrm{P(A,B)}=\frac{1}{2}\left(\frac{\mathrm{P(A)+P(B)}}{2}+\mathrm{P(A)P(B)}\right). \tag{6}$$

and for the three conjunct case:

$$\mathrm{P(A,B,C,)}=\frac{1}{2}\left(\frac{\mathrm{P(A)+P(B)+P(C)}}{3}+\mathrm{P(A)P(B)P(C)}\right) \tag{7}$$

For comparison purposes a simple averaging model was also fit. It calculated the mean of participants' component ratings.

### 6.1.2. Fitting the source reliability model to empirical data

Like many other accounts, the source reliability account views participants' judgments as expressions of posterior degrees of belief[8] (albeit in light of a report, not in light of the initial personality description). However, Eq. 2 requires priors for the calculation of the conjunction posterior. These priors for the component statements can be eliminated by rearranging Eq. 1 to derive the priors, and then using this expression to replace the priors in Eq. 2:

$$\mathrm{priorX} =(\alpha-\alpha\rho)/(((\rho+\alpha-\alpha\rho)/\mathrm{P(X|REP_X)})-\rho) \tag{8}$$

Model fitting then simply involved finding the prior reliability ($\rho$) that minimizes the sum of the squared deviations between the model's predicted conjunction ratings and participants' conjunction ratings.

For the fitting exercise, participants' ratings and the reliability parameter were constrained to lie within .000001 and .999999. We fit two variants of Bovens and Hartmann's model. The first model assumed that all participants (or all participants in each condition for Experiment 3) shared the same source reliability parameter (''Shared''). The second allowed one source reliability parameter per participant and condition (''Unique'').

Unlike the Bovens and Hartmann model, neither comparison model had any free parameters. As the comparison models turned out to fit the data better than the source reliability model, the source reliability model was not penalized for its free parameter. Model comparisons unfavorable toward the source reliability model are therefore arguably conservative.

### *6.2. Modeling results*

#### *6.2.1. Experiment 1*

As can be seen in Fig. 9, both ways of fitting the source reliability model resulted in underestimation of empirical conjunction ratings. The model that allowed a different value of the reliability parameter for each participant (unique model) produced a slightly better fit than did the model with a shared reliability parameter. Wyer's (1976) model produced the lowest sum of squared deviations. The data cluster around the identity line and approximately half of the predictions overestimate the conjunction and approximately half underestimate it. The simple averaging model tended to overestimate low conjunction ratings, while higher conjunction ratings cluster around the identity line. The fit of this model was only slightly better than that of the source reliability model.

#### *6.2.2. Experiment 3*

The model fits for Experiment 3 are summarized in Fig. 10 (scatter plots were similar to Experiment 1). As in Experiment 1, both comparison models better predicted participants' conjunction ratings than the source reliability model. This held even when each participant had an individual, unique reliability parameter. The achieved fits for unique and shared reliability parameters were yet again very similar. Thus, the ability of Bovens and Hartmann's model to predict the empirical data appears limited by the leverage of the reliability parameter as such. This is one of the issues we explore next in a general sensitivity analysis.

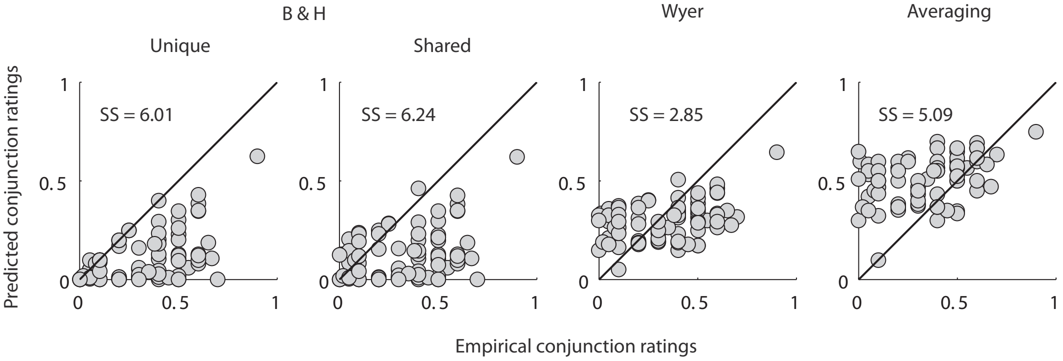

Fig. 9. The relationship between predicted conjunction ratings and actual conjunction ratings for each of the three models for Experiment 1. SS, sum of squared deviations. A lower SS value indicates a better fit (where 0 is a perfect fit).

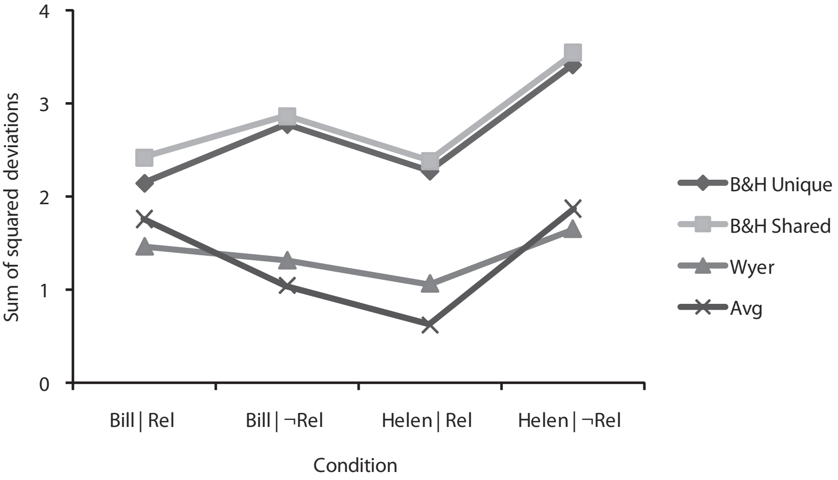

Fig. 10. Model fits for all scenario (Bill, Helen) and source reliability (reliable, unreliable) combinations. A lower sum of squared deviations value indicates a better fit (where 0 is a perfect fit).

### 6.3. Sensitivity analysis

We next asked a more general question. Could Bovens and Hartmann's model ever describe the empirical data? In other words, can the model produce posterior probabilities that match those supplied by human participants? We focused on the differences in ratings between P(L, U) and P(U) for those participants who committed the fallacy and asked whether the model could ever match the average empirical difference function (ΔP) for these participants. Among the participants who committed the fallacy, ΔP tended to be quite large. The median difference between ratings of P(L, U) and P(U) was .2 for the Bill scenario and .3 for the Helen scenario in Experiment 3. The median difference in Experiment 1 was .3.

Interestingly, Bovens and Hartmann's (2003) model does not allow for differences between posterior estimates of P(L, U) and P(U) that match the empirical differences in either experiment. The maximal positive difference (+ΔP MAX) allowed by the model is ~.17. In order to achieve this difference, the priors have to be set to ~1 (priorL) and ~.0001 (priorU) and the reliability parameter has to be set to ~.99. Thus, the L prior has to be near certain, the U prior near impossible, and the source has to be near fully reliable. Yet, even when the parameters are set as to maximize the difference, +ΔP MAX approaches but does not quite reach the empirical median +ΔP.

Thus far, we have only considered sources who, if thought to be unreliable, are perceived as unbiased. Given the noncommittal model, and given standard conjunction fallacy vignettes, this is arguably reasonable. Nevertheless, it may be informative to explore the model's ability to account for the data when $\alpha$ is free to vary, by itself or jointly, with $\rho$. Fig. 11 shows model fits for both experiments given the constraints used in the fits reported thus far ($\rho$ = free, $\alpha = .5$) as well as for these additional two cases.

As can be seen (Fig. 11), the fit improves if the randomization parameter ($\alpha$) is allowed to vary instead of the source reliability parameter ($\rho$). The model improves further, but only

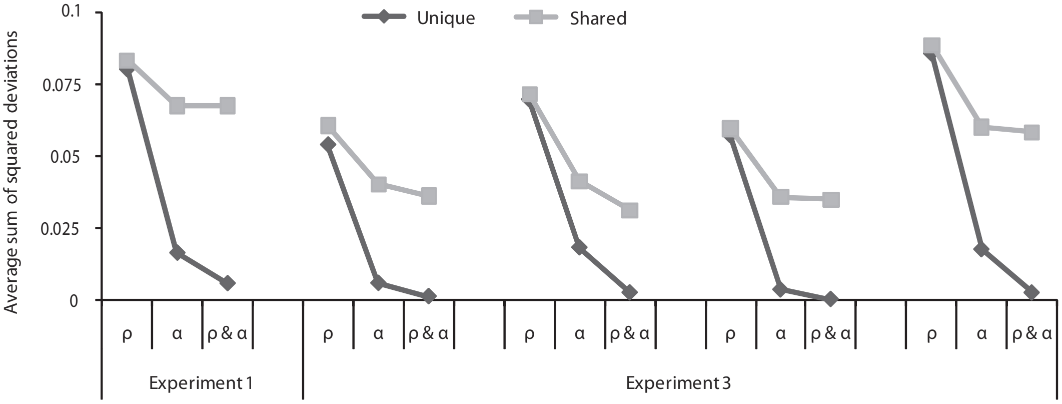

Fig. 11. Model fits as a function of which parameter(s) is (are) allowed to vary (e.g., $\rho$ means that the source reliability parameter was free to vary), and whether each participant has their own parameter (Unique) or whether the parameter is assumed to be identical across participants (Shared) for Experiments 1 and 3. The four separate plots for Experiment 3 correspond to each condition (in order: Bill Rel, Bill ¬Rel, Helen Rel, Helen ¬Rel). Average sum of squared deviations are presented instead of sum of squared deviations to enable comparisons across experiments (with different *N*).

for the unique fits, when both parameters are free to vary. Even with one free parameter per data point (for Experiment 3), the model cannot completely account for all data, as can be seen by the fit index being higher than zero (this is partly due to the model's inability to account for the following response pattern L = high probability, U = virtually impossible, and LU = medium probability). Noticeable is the fact that there is only a limited improvement in fit (if any) when both parameters are free to vary as opposed to when the randomization parameter is varied on its own. In fact, the randomization parameter seems sufficiently powerful to account for most of the variance on its own.

To demonstrate this, Fig. 12 shows model predictions for two component conjunction problems as a function of source reliability- and randomization-parameter values. As can be seen, source reliability ($\rho$) substantially influences $+\Delta P$ space (grey area) only for $\alpha = .5$ (middle column, Fig. 12). For more extreme $\alpha$ values, the randomization parameter almost completely determines the $\Delta P$ space. For low $\alpha$ values most of the area is grey and the fallacy should be committed; conversely, for high $\alpha$ values most of the area is white and the fallacy should not be committed. The model can thus, almost independently of the source reliability parameter, ''predict'' whether the fallacy should or should not be committed by setting the randomization parameter.

As the model, even with two free parameters, does not produce perfect fits (see Fig. 11), it is clear that the model cannot predict all possible posterior LU ratings given L and U. Nevertheless, when the randomization parameter ($\alpha$) is also free to vary, the model is arguably too unconstrained. In particular, the model could readily fit data that seem nonsensical in the context of conjunction problems. For example, the model could fit two individual statements of low probability—P(1) = .01, P(2) = .01—whose conjunct, however, is viewed as very probable, P(1, 2) = .99 (with a sum of squared deviation of <1e-4). The model

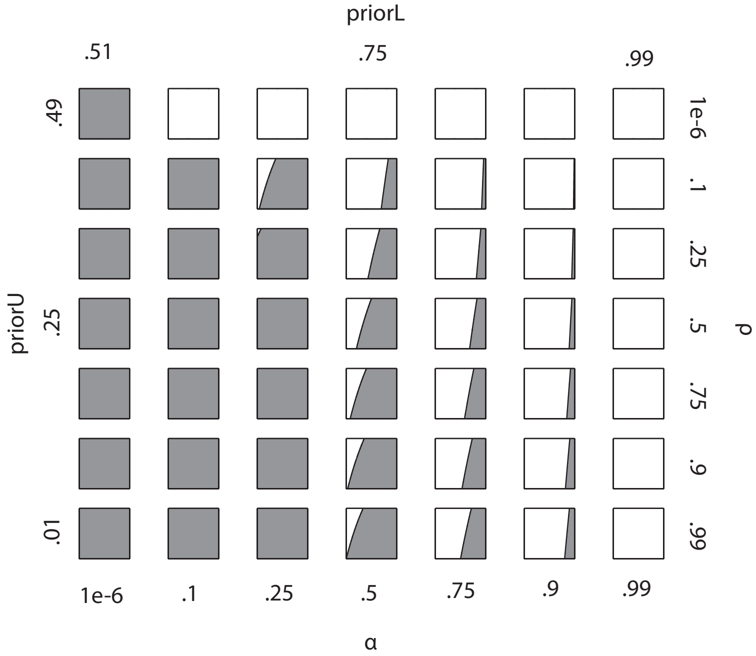

Fig. 12. Each square shows positive (grey) and negative (white) ΔP space as a function of the prior probability of L and U, for given $\rho$ and $\alpha$ parameter values (.99 = .999999).

achieves the good fit by making the source highly likely to be unreliable and highly likely to produce a negative report if unreliable ($\rho < .01$, $\alpha < .01$). To see why this seems problematic, a verbal example is useful. The hypothetical ratings might, for example, correspond to the probability that ''Linda is a republican'' and that ''Linda is an NRA (National Rifle Association) member,'' two claims which seem very unlikely given her description, while their conjunction—''Linda is a republican and is an NRA member''—would nevertheless be very likely. It seems difficult to see such ratings as plausible.

If a model is so powerful that it can equally fit both data that one would want to fit and data that one would not, then the mere fact of a good fit seems uninformative, and the model no longer seems explanatory (see also Roberts & Pashler, 2000). Nor is it clear that values of the randomization parameter ($\alpha$) other than .5 have intuitively desirable consequences once a sufficiently broad range of possible cases is considered. For these reasons, treating $\alpha$ as a free parameter seems undesirable. In order to model not just source reliability but also bias, it would seem that a more sophisticated conception of partially reliable sources is required.

## 7. General discussion

The present study set out to test a source reliability account (Bovens & Hartmann, 2003) of the conjunction fallacy. To this end, three experiments were conducted. Two of the experiments assessed predictions relating to changes in the frequency of the conjunction fallacy. Neither experiment was able to confirm the predictions of Bovens and Hartmann's model. Instead, some evidence against the model was found in both Experiment 1 and 3. The failure to detect a difference in rates in Experiment 1 may have been due to the failure of participants to infer source reliability from the probability of statements, rather than a failure of the model per se. However, Experiment 2 provided evidence against this possibility. The inferred source reliability matched the model's predictions (i.e., L > LU > U). In summary, then, neither the addition of one additional (likely or unlikely) component to standard problems, nor a change in the reliability of the source that provides the statements, appears to affect the incidence of the conjunction fallacy.

The model fitting results agreed well with the experimental results. The source reliability model fit the data less well than Wyer's (1976) model and less well than a simple averaging model. Further exploration showed that the ability of source reliability to affect predicted conjunction ratings is severely limited given an unbiased source. However, the model's ability to fit the data was dramatically improved by fitting the randomization parameter instead of the source reliability parameter. It was argued that the most parsimonious application of the model is one which fixes the randomization parameter (and specifically assumes no bias). In this case, three experiments, model fitting and a sensitivity analysis provide very little support for Bovens and Hartmann's model.

The present analyses do, of course, not rule out *other* source reliability accounts. As noted initially, the tested model conceptualizes source reliability and its influence on reports in a very simple manner. In our view, the most interesting aspect of the model is that it shows that the fallacy may indeed be the normative response *even* when the influence of source reliability is modeled in a maximally noncommittal way. Hence, the simplicity of the model is arguably as much a strength as it is a weakness.

Our empirical results do, however, suggest that greater psychological plausibility in the modeling of source reliability seems necessary if source reliability is to provide an adequate explanation of participants' behavior. It is encouraging then that the probabilistic framework used by Bovens and Hartmann is so readily extendable. Hartmann and Meijs (2009) have, for example, recently proposed a similar model that allows the original scenario description to come from a partially reliable source. They also model the reports of unreliable sources as dependent on the prior probability of facts, rather than being determined by a randomization parameter.

Thus, it is conceivable, that more sophisticated, or psychologically more plausible, source reliability models will prove descriptively valid. Nevertheless, until such models are applied to the conjunction fallacy, *and* are evaluated empirically, it seems prudent to conclude (be it temporarily) that source reliability does not explain the conjunction fallacy.

If the conjunction fallacy is not caused by source reliability effects, then what causes it? It has recently been argued that an erroneous combination of component probabilities is the key (Nilsson, 2008), or at least a major (Wedell & Moro, 2008) component of the conjunction fallacy. Given that both a simple averaging model and Wyer's (1976) model produced better fits than the source reliability account, it is tempting to conclude that participants erroneously implement some type of averaging strategy when evaluating compound probabilities. It is, however, our belief that such a conclusion is also premature.

Simple averaging models predict that everyone who rates component statements ever so slightly differently will commit the fallacy. Given that not everyone commits the fallacy (see e.g., Tversky & Kahneman, 1983), not everyone uses an averaging strategy. However, if one weight (or noise) parameter per component rating is allowed, as in, for example, Oden and Anderson's (1971) model (see Fantino, Kulik, Stolarz-Fantino, & Wright, 1997 for a qualitative assessment), or as in Birnbaum, Anderson, and Hynan (1990), then given ''correct'' parameters, averaging models can account both for fallacious and nonfallacious responding. Moreover, if one weight per estimate is allowed, and the weights are otherwise unconstrained, it seems in principle possible to account for *any* conjunction rating given two-component estimates. It may be argued that this flexibility is unproblematic as one can estimate parameters using group data, thereby increasing the number of data points per free parameter. However, this argument is questionable as long as it is unclear what psychological meaning the free parameters carry. Arguing that they signify the importance of individual components (e.g., Fantino et al., 1997, p. 99) seems circular if the importance (i.e., weight) follows solely from the fitting procedure itself (i.e., they are ''important'' in explaining the data given the computational model chosen). Without meaningful interpretations, noise-, or weight-, parameters run the risk of remaining post-hoc additions included to improve model fits.

Unlike an averaging model without free parameters, Wyer's (1976) model predicts that if component probabilities differ very little, fallacies will not be committed. Thus, it can predict nonfallacies. This was presumably one of the reasons why the model outperformed the other models in our tests. However, as some of our participants rated one component as likely, and rated the other as unlikely, and yet did not commit the fallacy, it too cannot describe all the data.

Representativeness does not appear to explain the conjunction fallacy completely either. Whether the statement ''Linda is a bank teller and is active in the feminist movement'' is representative of Linda is arguably independent of the reliability of the source making the statement. This supposed independence is not reflected in ratings of statements from sources that vary in reliability. People report lower ratings for statements by unreliable source compared to statements by reliable sources (Experiment 3). The representativeness account also seems to lack a mechanism by which source reliability can be extracted from the prior likelihood of statements (Experiment 2).

We are then returned to the position noted in the introduction that no present theory seems able to account for all factors that affect the fallacy.[9] While present theories of the conjunction fallacy all explain why many participants commit the fallacy (e.g., representativeness), or why fewer participants commit the fallacy in some contexts (e.g., natural frequency representations), none of these theories individually seems to explain why some participants do not commit the fallacy, or why others continue to commit the fallacy despite being offered, for example, ''ecologically valid'' stimuli.

The fact that the conjunction fallacy is sensitive to contextual effects and individual differences suggests that current approaches collectively may be less than ideal. Fitting models to between-subject data, when qualitative differences in behavior exist, may result in a good description of the behavior of the group, but a poor description of the behavior of any one individual in that group (Gallistel, Fairhurst, & Balsam, 2004). If it is true that participants use different strategies, the cause of the conjunction fallacy will probably not be found by analyses that ignore those differences and treat them simply as noise.

In summary, the empirical predictions derived from Bovens and Hartmann's (2003) model of the conjunction fallacy were not confirmed. Instead some evidence against these predictions was found. Modeling and sensitivity analyses showed that the model either cannot predict empirical data (when $\rho$ is a free parameter) or can predict almost any data (when both $\rho$ and $\alpha$ are free to vary). The latter case is arguably uninformative and relies on manipulating a parameter, which if set to anything but .5 lacks clear meaning in the context of conjunction fallacy problems. Nevertheless, it is conceivable that Bovens and Hartmann's account is, in principle correct, or is at least part of the solution to the problem, but that a differently specified source reliability model is required. Even as things stand, however, the idea that source reliability *should* impact inferences, and decisions based on those inferences, deserves serious consideration and opens up avenues for reevaluating other apparent violations of rationality.

## Notes

1. A Google Scholar search with the term ''conjunction fallacy'' OR ''conjunction effect'' yielded 2400 hits (18.07.2009).
2. Stolarz-Fantino, Fantino, and Kulik (1996) found an effect after description removal. However, they provided participants with the component probabilities. The purpose of the description is to set the prior probabilities for the components. Thus, although the description was absent, its function was fulfilled. Similarly, Hertwig and Gigerenzer (1999) found that no participants committed the fallacy when presented with a ''frequency'' scenario, presented as if it were an opinion poll (Study 3). However, they failed to replicate the ''no fallacy'' finding (Study 4) and to our knowledge it has not been replicated since.
3. Bovens and Hartmann's (2003) exposition is somewhat ambiguous with regard to what the negation of a report (i.e., $\neg REP_X$) is intended to mean in the context of the conjunction fallacy. In other, preceding examples (Chapter 3 and 4), the negation implies the absence of a report. However, the reported equations (p. 140) suggest that the negation of a report ($\neg REP_X$) is best viewed as a negative report. The posterior probability of a single belief given a single report to that belief is described as $P(X|REP_X)$—as in Eq. 1 here. If, $\neg REP_Y$ were the absence of a report to Y, then $P(X|REP_X, \neg REP_Y)$ would arguably be the appropriate term for the report of X and the absence of the report Y. Consequently, we adopt the interpretation that is required for a consistent mapping between the network (p. 86) and the derived equations (p. 140). We assume that $\neg$REP stands for a negative report, and that the absence of a report is equivalent to not activating the corresponding node in the Bayesian Network (see Fig. 1B here). This interpretation, results in the posterior probability of a single statement being equivalent whether computed using Eq. 1, or whether by an implementation of the network (Fig. 1B).

4. It might seem odd to represent conjunction problem statements as binary variables. Statements can either be made, not made or statements can be negated. Thus, a discrete variable with three states might seem appropriate. However, if the absence of a report is uninformative with regard to the prior probability of the underlying belief or the reliability of the source, these two representations produce the same posterior belief upon a positive report. That is, observing $\mathrm{REP}_\mathrm{L}$ in Fig. 1B causes the same belief updating, as observing $\mathrm{REP}_\mathrm{L}$ & $\mathrm{NoREP}_\mathrm{U}$ in a network where NoREP is represented explicitly (i.e., REP variables with three states). We have maintained Bovens and Hartmann's (2003) binary representation.
5. The crucial result of the model can also arise without this independence assumption. For example, the result still holds for a model in which L is less likely under U than under ¬U.
6. Reported in Jarvstad and Hahn (2009).
7. ''Quite reliable'' was used to maintain the natural language description in the original problems. The qualifier ''quite'' was used to avoid ceiling and floor effects.
8. There appears to be no straightforward way to assess priors empirically in classical conjunction problems given Bovens and Hartmann's (2003) interpretation.
9. Although Bovens and Hartmann's model, too, is silent on many of the factors that moderate the fallacy, it would presumably be straightforward to extend it, given the general probabilistic framework it adopts, to account for many of these. An increase in the rate of the fallacy, when people rank rather than rate statements, could, for example, be modeled by a random response for ties.

## Acknowledgments

We thank our action editor Michael Lee, and Luc Bovens, David Lagnado, and an anonymous reviewer, for helpful suggestions and comments. Andreas Jarvstad was supported by the Economic and Social Research Council (ESRC), UK.

## References

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

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

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

Birnbaum, M. H., Anderson, C. J., & Hynan, L. G. (1990). Theories of bias in probability judgment. In J. P. Caverni, J. M. Fabre, & M. Gonzalez (Eds.), *Cognitive biases* (pp. 477–498). Oxford, England: North-Holland.

Birnbaum, M. H., & Stegner, S. E. (1979). Source credibility in social judgment: Bias, expertise and the judge's point of view. *Journal of Personality and Social Psychology*, *37*, 48–74.

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford, England: Clarendon Press.

Chaiken, S., & Maheswaran, D. (1994). Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgment. *Journal of Personality and Social Psychology*, *66*, 460–473.

Corner, A. J., Harris, A. J. L., & Hahn, U. (2010). Conservatism in belief revision and participant skepticism. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (pp. 1625–1630). Austin, TX: Cognitive Science Society.

Crupi, V., Fitelson, B., & Tentori, K. (2008). Probability, confirmation, and the conjunction fallacy. *Thinking and Reasoning*, *14*, 182–199.

Donovan, S., & Epstein, S. (1997). The difficulty of the Linda conjunction problem can be attributed to its simultaneous concrete and abstract representation, and not to conversational implicature. *Journal of Experimental Social Psychology*, *33*, 1–20.

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

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Orlando, FL: Harcourt Brace.

Epstein, S., Denesraj, V., & Pacini, R. (1995). The Linda problem revisited from the perspective of cognitive-experiential self-theory. *Personality and Social Psychology Bulletin*, *21*, 1124–1138.

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

Fisk, J. E. (2002). Judgments under uncertainty: Representativeness or potential surprise? *British Journal of Psychology*, *93*, 431–449.

Fisk, J. E. (2004). Conjunction fallacy. In R. F. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgment, and memory* (pp. 23–42). London: Psychology Press.

Fisk, J. E., & Pidgeon, N. (1996). Component probabilities and the conjunction fallacy: Resolving signed summation and the low component model in a contingent approach. *Acta Psychologica*, *94*, 1–20.

Fisk, J. E., & Pidgeon, N. (1998). Conditional probabilities, potential surprise, and the conjunction fallacy. *Quarterly Journal of Experimental Psychology*, *51A*, 655–681.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439–453.

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 13124–13131.

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

Hahn, U., Harris, A. J. L., & Corner, A. J. (2009). Argument content and argument source: An exploration. *Informal Logic*, *29*, 337–367.

Harris, A. J. L., & Hahn, U. (2009). Bayesian rationality in evaluating multiple testimonies: Incorporating the role of coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1366–1373.

Hartmann, S., & Meijs, W. (2009). Walter the banker: The conjunction fallacy reconsidered. *Synthese*. DOI 10.1007/s11229-009-9694-6.

Hertwig, R., Benz, B., & Krauss, S. (2008). The conjunction fallacy and the many meanings of and. *Cognition*, *108*, 740–753.

Hertwig, R., & Gigerenzer, G. (1999). The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, *12*, 275–305.

Jarvstad, A., & Hahn, U. (2009). Unreliable sources and the conjunction fallacy. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society* (pp. 3034–3039). Austin, TX: Cognitive Science Society.

Jeffreys, H. (1961). *Theory of probability* (3rd Ed.). Oxford, England: Oxford University Press, Clarendon Press.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). New York: Cambridge University Press.

Kahneman, D., & Tversky, A. (1982). Subjective probability: A judgment of representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 33–47). Cambridge, England: Cambridge University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kruglanski, A. W., & Thompson, E. P. (1999). Persuasion by a single route: A view from the unimodel. *Psychological Inquiry*, *10*, 83–109.

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*, 662–668.

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

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

McKenzie, C. R. M., Wixted, J. T., & Noelle, D. C. (2004). Explaining purportedly irrational behavior by modeling scepticism in task parameters: An example examining confidence in forced-choice tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 947–959.

Nilsson, H. (2008). Exploring the conjunction fallacy within a category learning framework. *Journal of Behavioral Decision Making*, *21*, 471–490.

Oden, G. C., & Anderson, N. H. (1971). Differential weighting in integration theory. *Journal of Experimental Psychology*, *89*, 152–161.

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology*, *34*, 243–281.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Schum, D. A. (1981). Sorting out the effects of witness sensitivity and response-criterion placement upon the inferential value of testimonial evidence. *Organizational Behavior and Human Performance*, *27*, 153–196.

Sides, A., Osherson, D., Bonini, N., & Viale, R. (2002). On the reality of the conjunction fallacy. *Memory & Cognition*, *30*, 191–198.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*, 161–188.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*, 672–695.

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.

Stolarz-Fantino, S., Fantino, E., Zizzo, D. J., & Wen, J. (2003). The conjunction effect: New evidence for robustness. *American Journal of Psychology*, *116*, 15–34.

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

Teigen, K. H., Martinussen, M., & Lund, T. (1996b). Conjunction errors in the prediction of referendum outcomes: Effects of attitude and realism. *Acta Psychologica*, *93*, 91–105.

Tentori, K., Bonini, N., & Osherson, D. (2004). The conjunction fallacy: A misunderstanding about conjunction? *Cognitive Science*, *28*, 467–477.

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, England: Cambridge University Press.

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

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.

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

Wyer Jr., R. S. (1970). The quantitative prediction of belief and opinion change: A further test of a subjective probability model. *Journal of Personality and Social Psychology*, *16*, 559–571.

Wyer Jr., R. S. (1976). An investigation of the relations among probability estimates. *Organizational Behavior and Human Performance*, *15*, 1–18.