Cognition 113 (2009) 26–36

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

On the provenance of judgments of conditional probability

Jiaying Zhao *, Anuj Shah, Daniel Osherson

Department of Psychology, Green Hall, Princeton, NJ 08540, United States

ARTICLE INFO

Article history:
Received 17 February 2009
Revised 9 July 2009
Accepted 10 July 2009

Keywords:
Conditional probability
Reasoning
Judgment

ABSTRACT

In standard treatments of probability, $Pr(A|B)$ is defined as the ratio of $Pr(A \cap B)$ to $Pr(B)$, provided that $Pr(B) > 0$. This account of conditional probability suggests a psychological question, namely, whether estimates of $Pr(A|B)$ arise in the mind via implicit calculation of $Pr(A \cap B)/Pr(B)$. We tested this hypothesis (Experiment 1) by presenting brief visual scenes composed of forms, and collecting estimates of relevant probabilities. Direct estimates of conditional probability were not well predicted by $Pr(A \cap B)/Pr(B)$. Direct estimates were also closer to the objective probabilities defined by the stimuli, compared to estimates computed from the foregoing ratio. The hypothesis that $Pr(A|B)$ arises from the ratio $Pr(A \cap B)/[Pr(A \cap B) + Pr(\overline{A} \cap B)]$ fared better (Experiment 2). In a third experiment, the same hypotheses were evaluated in the context of subjective estimates of the chance of future events.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Consider the chance that France wins the next World Soccer Cup (F). Now consider it again but this time assuming that Italy is eliminated before the quarter finals (I). The latter judgment is your *conditional probability* for *F given I*, denoted $Pr(F|I)$. How does the mind estimate such chances?

The matter is central to Bayesian accounts of updating a probability distribution Pr to accommodate the information that an event B has occurred (for sure). According to Bayesians (Hacking, 2001), the revised distribution should be $Pr(\cdot|B)$, which assigns a given event A the conditional probability $Pr(A|B)$. The psychology of updating has typically been investigated in settings that offer *base rates* and *likelihoods*, thereby authorizing use of Bayes' Theorem to compute conditional probability. The resulting literature considers whether these latter quantities are suitably deployed (Koehler, 1996; Tversky & Kahneman, 1982). When Bayes' Theorem is conceptualized as a means of calculating conditional probability, however, a different question is highlighted: what does conditional probability represent to the reasoner, in other words, how is it mentally defined?

This question is also connected to inductive inference. For, the conditional probability of A given B is a plausible interpretation of the "strength" of an argument with premise B and conclusion A. Inductive inference has been extensively examined in these terms (Feeney & Heit, 2007). Such studies illuminate the conditions affecting estimates of conditional probability, e.g., typicality (Murphy & Ross, 2005), the kind and number of categories involved (Ross et al., 1996, 1999), and the role of similarity (Weber & Osherson, in press). But these studies provide limited information about the interpretation of conditional probability in the mind of the reasoner. This is because they rely on conditional probability to measure inductive strength but do not investigate how it is mentally represented.

Fox and Levav (2004) offer insightful analysis of conditional probability assessments in a setting that allows counting of target events. They examine influences on the categories participants' count, and the use made of these numbers. Many situations, however, do not lend themselves to such numerical strategies, either because

* Corresponding author.
E-mail address: jiayingz@princeton.edu (J. Zhao).

0010-0277/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2009.07.006

there are too many instances of the relevant categories (e.g., common social situations), or too few (as in the soccer example above). Accordingly, we here examine how people estimate conditional probabilities when it is difficult (Experiments 1 and 2) or impossible (Experiment 3) to count events.

Let us set aside two ideas about the provenance of judgments of conditional probability. First, we take the meaning of $Pr(A|B)$ to be an issue distinct from the interpretation of conditionals like "If B then A" unless it is conjectured that evaluation of $Pr(A|B)$ proceeds via $Pr(\text{If } B \text{ then } A)$. No such conjecture informs the present discussion for reasons given at the end. Nor is it helpful to define conditional probability in terms of Bayes' Theorem

$$Pr(A|B) = \frac{Pr(B|A) \times Pr(A)}{Pr(B)}$$

inasmuch as conditional probability appears on both sides of the equation. More generally, the proposal that $Pr(A|B)$ arises from the probability of A in the distribution that results from updating in light of B risks circularity if it is left open that updating proceeds via conditionalization.

Axiomatic presentations of probability (as in Ross (1988)) typically define conditional probability from absolute probability via the equation

$$Pr(A|B) =_{\text{def}} \frac{Pr(A \cap B)}{Pr(B)} \quad \text{provided } Pr(B) > 0. \tag{1}$$

From a normative perspective, the equation can be justified independently of human psychology, for example, in terms of nondominated quadratic penalties (Bernardo & Smith (1994, p. 89)), or fair betting rates (Jeffrey, 2004). Our interest in (1) is descriptive, however, rather than normative. At a descriptive level, the definition invites the hypothesis that judgments of conditional probability arise by implicit calculation of the ratio of the two absolute probabilities shown above. We intend this hypothesis in the following sense. When a reasoner is confronted with a request for $Pr(A|B)$, she computes $Pr(A \cap B)$ and $Pr(B)$ just as if she were asked for each of the latter probabilities separately, then divides. Let RH denote the hypothesis that judgments of conditional probability arise in this way from implicit calculation of the ratio shown in (1).

There are at least three reasons to doubt the plausibility of RH. First, according to (1), $Pr(A|B) = Pr(B|A)$ only if $Pr(A) = Pr(B)$. Yet the inversion of conditional probabilities is a common feature of judgment even when it is recognized that $Pr(A) \neq Pr(B)$ (Dawes, Mirels, Gold, & Donahue, 1993; Eddy, 1982). Such inversion undermines the conviction that most people understand the concept of conditional probability. On the other hand, conditional inversion is not prevalent in the experiments reported below.

Another reason to doubt RH is that it conflicts with intuition in cases involving continuous sample spaces (Hájek, 2003). For example, suppose a number is drawn uniform randomly from $[0, 1]$, and let $B = \{.6, .7, .8\}$.[1] It seems that the chance of falling below.75 given that a member of B is drawn equals 2/3 whereas (1) recognizes no such conditional probability because $Pr(B) = 0$. Examples of this character have prompted axiomatizations that reverse the roles of conditional and absolute probability. In Popper (1959), for example, conditional probability is primitive and $Pr(A)$ is defined as $Pr(A|\Omega)$ where Ω is the certain event. This objection will be mitigated in the present study by limiting attention to conditional probabilities whose conditioning events are likely to have positive (subjective) probability.

The third and most important reason for scepticism about RH relates to the logical forms of numerator and denominator in (1). Specifically, $A \cap B$ is an element of the binary partition $(A \cap B) \cup (\overline{A} \cap B)$ of B. Because each element exhibits greater specificity than B (by virtue of its intersection with A or $\overline{A}$), their probabilities may be overestimated with respect to that of B, yielding *subadditivity*: $Pr(B) < Pr(A \cap B) + Pr(\overline{A} \cap B)$. Subadditive judgment was originally documented in Fischhoff, Slovic, and Lichtenstein (1978) and confirmed by Russo and Kolzow (1994), Tversky and Koehler (1994), and others. It is the principal motivation for *Support Theory* (Rottenstreich & Tversky, 1997, 1999; Tversky & Koehler, 1994), which exhibits the judged probability of an event as a function of the evidential support brought to mind by the event's description. In sum, events that are mentally represented as intersections may recruit additional support, exaggerating their probability in comparison with events represented more simply. Indeed, such recruitment may be one reason for the *conjunction fallacy*, in which estimates of $Pr(A \cap B)$ exceed those of $Pr(B)$ (see Tentori, Bonini, & Osherson, 2004; Tversky & Kahneman, 1983 and references cited there). In comparison, the expression of conditional probability involves only single events, with no explicit intersections. A bias may thus affect the numerator of (1) but not be present in judgments of $Pr(A|B)$. Such a situation would lead RH to overestimate conditional probabilities.

The latter argument against RH may not be decisive, however. For one thing, it begs the question to claim that event-intersections are absent from the mental representation of $Pr(A|B)$, for they are present if conditional probabilities are interpreted as the ratio in (1). The matter seems to depend on how "implicit" the mental ratio posited by RH is supposed to be. Moreover, partition elements sometimes attract a dearth of support rather than a surplus, leading to *superadditivity*. For example, in Macchi, Osherson, and Krantz (1999), one group of undergraduates (in Italy) gave their probability that the Duomo in Milan is taller than Notre Dame in Paris, whereas another group gave the probability that Notre Dame is taller than the Duomo. All participants were informed that the heights are different. Despite many responses of 0.5, the sum of the average answers for the two groups was only 0.72 (answers of 0.5 may be expressions of ignorance, Fischhoff & Bruine de Bruin (1999)). Superadditive judgment is also reported in Cohen, Dearnaley, and Hansel (1956) Sloman, Rottenstreich, Wisniewski, Hadjichristidis, and Fox (2004). Brenner et al. (1999) document superadditivity for binary partitions in the case in which one of the partition elements is itself a disjunction. See Idson, Krantz, Osherson, and Bonini (2001) for experimental test of a revision of Support Theory that is consistent with both sub- and superadditivity. In the present context,

[1] The uniform distribution over [0,1] sets the probability of sampling an interval $I \subseteq [0, 1]$ equal to the length of I. Single points in $[0, 1]$ thus have zero probability (since they represent intervals of length zero).

the point is that (revised) Support Theory does not predict unambiguously that the binary partition $(A \cap B) \cup (\overline{A} \cap B)$ of B will lead to subadditivity, and overestimation of $Pr(A \cap B)$. (But in fact, Experiments 2 and 3 below will show considerable subadditivity.)

The *a priori* plausibility of RH is thus open to debate. The foregoing discussion nonetheless motivates a second hypothesis, in which all events have the same logical form. Since $Pr(B) = Pr(A \cap B) + Pr(\overline{A} \cap B)$, (1) implies:

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(A \cap B) + Pr(\overline{A} \cap B)} \quad \text{provided } Pr(B) > 0. \tag{2}$$

Let RH′ be the hypothesis that judgments of conditional probability arise from implicit calculation of the ratio shown in (2). Mathematically, (1) and (2) are equivalent. It will be seen, however, that RH′ enjoys considerably greater accuracy than RH as a psychological hypothesis, perhaps because its numerator and denominator rely equally on event intersections.[2]

The first two experiments reported here involve brief visual presentation of forms of varying shape, color, and position. To formulate predictions associated with RH and RH′ in this context, we rely on the following notation. Let two perceptual categories A, B be given (e.g., red, square), and let **S** be a specified visual scene. Then, for a given experimental participant:

$Pr[\text{dir}](B)$ denotes the judged probability that a form drawn randomly from **S** is B, and likewise for $Pr[\text{dir}](A \cap B)$ and $Pr[\text{dir}](\overline{A} \cap B)$. $Pr[\text{dir}](A|B)$ denotes the judged probability that such a form is A assuming that it is B ("dir" stands for "direct").

$Pr[\text{ind}](A|B)$ denotes the ratio of $Pr[\text{dir}](A \cap B)$ to $Pr[\text{dir}](B)$. Thus, $Pr[\text{ind}](A|B)$ is the conditional probability of A given B as computed from (1) ("ind" stands for "indirect").

$Pr[\text{ind}'](A|B)$ denotes $Pr[\text{dir}](A \cap B)$ divided by $Pr[\text{dir}](A \cap B) + Pr[\text{dir}](\overline{A} \cap B)$. Thus, $Pr[\text{ind}](A|B)$ is the conditional probability of A given B as computed from (2).

$Pr[\text{obj}](A|B)$ denotes the percentage of A's in **S** among the B's in **S**, i.e., the true conditional probability in **S** of A assuming B – and similarly for $Pr[\text{obj}](B), Pr[\text{obj}](A \cap B)$ and $Pr[\text{obj}](\overline{A} \cap B)$ ("obj" stands for "objective").

We understand RH and RH′ to respectively entail:

RH *predictions*: (3)

(a) $Pr[\text{ind}](A|B)$ is an unbiased estimate of $Pr[\text{dir}](A|B)$.

(b) $Pr[\text{dir}](A|B)$ and $Pr[\text{ind}](A|B)$ are equally close to $Pr[\text{obj}](A|B)$.

RH′ *predictions*: (4)

(a) $Pr[\text{ind}'](A|B)$ is an unbiased estimate of $Pr[\text{dir}](A|B)$.

(b) $Pr[\text{dir}](A|B)$ and $Pr[\text{ind}'](A|B)$ are equally close to $Pr[\text{obj}](A|B)$.

These predictions will be tested in the three experiments described below.

In the extensional setting of Experiments 1 and 2, $Pr[\text{dir}](A|B)$ was elicited via two kinds of wording. In the probability condition, participants were asked a question of the form: "Suppose that a B is chosen at random from the array; what is the probability that it is an A?" The frequency version of this question was: "What percent of the B's in the array are A's?" Similar wordings were used for $Pr[\text{dir}](A \cap B)$ and $Pr[\text{dir}](B)$. The two formulations test the robustness of our results inasmuch as frequency formats sometimes yield estimates more consistent with the probability calculus (Fiedler, 1988; Mellers, Hertwig, & Kahneman, 2001; Tversky & Kahneman, 1983). In the present experiments, the impact of alternative formats was minimal. The third experiment employed an intensional setting where participants estimated the probability of future, singular events.

2. Experiment 1

The primary purpose of the first experiment was to test RH through its predictions (3). Our design allowed the same participant to be queried, without knowing it, about $Pr(A|B), Pr(A \cap B)$, and $Pr(B)$ for the same events A, B. This allowed within-subject comparisons without intervention of the participant's own theory of conditional probability.

2.1. Participants

Forty-five undergraduate students from Princeton University participated in exchange for partial course credit (32 female, mean age 20.09 years, SD = 1.02).

2.2. Materials

Participants viewed 12 sets of geometric shapes on a computer screen. Each set was a mixture of 20 triangles, squares, and circles in blue, red, and green (all three shapes and all three colors appeared in every matrix). A given set was shown four times, with each display lasting one second. The shapes in a given display were arrayed as a 4×5 matrix, their respective positions individually randomized for each presentation. Fig. 1 illustrates one display for one of the 12 sets. The purpose of multiple brief, randomized displays of a given set was to prevent responses based on counting.

The four displays of a given set were initiated by a "Ready" button controlled by the participant. Henceforth, by a trial associated with a given set is meant the successive display of its four randomized matrixes.

For each set we chose one color and one shape to serve as the categories A and B evoked in the Introduction. A different choice was made for each of the 12 sets; for six sets

[2] Let us observe a structural advantage of RH′ compared to RH. Eq. (2) confines $Pr[\text{dir}](A|B)$ to the unit interval whereas (1) allows numbers beyond 1. Thus, RH′ but not RH is guaranteed to produce numbers that look like probabilities.

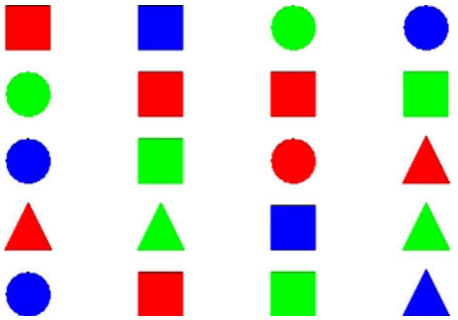

Fig. 1. Presentation of a stimulus set in Experiments 1 and 2. In the actual experiments the shapes were colored in blue, red, and green. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Objective probabilities in the sets of stimuli used in Experiments 1 and 2.

Level	$Pr\,[\text{obj}]\,(B)$	$Pr\,[\text{obj}]\,(A \cap B)$	$Pr\,[\text{obj}]\,(\overline{A} \cap B)$	$Pr\,[\text{obj}]\,(A\|B)$
Low	0.3	0.1	0.2	0.33
Medium	0.6	0.4	0.2	0.67
High	0.9	0.8	0.1	0.89

A was a color and B a shape, the reverse held for the other six. The sets were designed so that $Pr\,[\text{obj}]\,(A \cap B)$ and $Pr\,[\text{obj}]\,(B)$ were either .1 and .3, .4 and .6, or .8 and .9. These three cases yield $Pr\,[\text{obj}]\,(A|B)$ equal to .33, .67, or .89, respectively. Four sets fell into each of these cases, called low, medium, and high levels in what follows. Table 1 summarizes the objective probabilities figuring in the experiment.

2.3. Procedure

Each participant served in both the probability and frequency conditions (the order was counterbalanced). In each condition, the participant viewed the 12 sets three times, once for each query $Pr\,(B), Pr\,(A \cap B)$, or $Pr\,(A|B)$. The colors and shapes representing A and B were the same in the three trials for a given set. The 36 resulting trials were presented in individualized random order under the constraint that a given set not appear twice in a row. Following each trial, the participant responded to one question corresponding to $Pr\,(B), Pr\,(A \cap B)$, or $Pr\,(A|B)$. For the probability condition, the questions are illustrated as follows.

Sample probability questions:

$Pr\,(B)$	What is the probability that a randomly selected shape in the set is red?
$Pr\,(A \cap B)$	What is the probability that a randomly selected shape in the set is a red square?
$Pr\,(A\|B)$	What is the probability that a randomly selected shape in the set is square assuming that it is red?

Table 2
Average direct and indirect estimates segregating by level from Experiment 1 (standard deviations in parentheses).

Level	B	$A \cap B$	$A\|B$	ind $(A\|B)$
Low	0.31 (0.07)	0.19 (0.08)	0.35 (0.10)	0.70 (0.32)
Medium	0.58 (0.07)	0.50 (0.10)	0.63 (0.10)	0.93 (0.22)
High	0.86 (0.04)	0.80 (0.07)	0.82 (0.12)	1.03 (0.35)

For the frequency condition, the corresponding questions were:

Sample frequency questions:

$Pr\,(B)$	What percent of the shapes in the set are blue?
$Pr\,(A \cap B)$	What percent of the shapes in the set are blue circles?
$Pr\,(A\|B)$	What percent of the blue shapes in the set are circles?

Thus, in both conditions, a given set yielded values for each of $Pr\,[\text{dir}]\,(B), Pr\,[\text{dir}]\,(A \cap B)$ and $Pr\,[\text{dir}]\,(A|B)$. Participants entered their answers using either decimals, fractions, or percents according to their preference. The experiment began with explanation of the task, followed by practice trials. Participants were not informed that sets would be repeated (with different queries); none seem to have discovered this fact. Between the two conditions (probability and frequency), participants completed a 5 min distraction task.

2.4. Results

2.4.1. Average responses

The probability and frequency conditions produced very similar numbers; across all participants, the average discrepancy between responses to corresponding queries was only 0.02. For each participant, we averaged their 12 values of $Pr\,[\text{dir}]\,(B)$ for probability and for frequency responses; we then performed a paired t-test on these numbers across the 45 participants, with nonsignificant result ($p > .05$). The same is true for $Pr\,[\text{dir}]\,(A \cap B)$ and $Pr\,[\text{dir}]\,(A|B)$. The two conditions were therefore collapsed; each judgment was taken to be the mean of the responses to its probability and frequency variants. For a given participant, we averaged the response to each query – $Pr\,[\text{dir}]\,(B), Pr\,[\text{dir}]\,(A \cap B)$ or $Pr\,[\text{dir}]\,(A|B)$ – at each level (low, medium, high). Each of these nine categories of numbers (three levels by three queries) was then averaged across the 45 participants. To compute $Pr\,[\text{ind}]\,(A|B)$, for each participant and each set, we divided her estimate of $Pr\,[\text{dir}]\,(A \cap B)$ by her estimate of $Pr\,[\text{dir}]\,(B)$, for the events A, B specific to that participant and stimulus. The results are shown in Table 2.

2.5. Test of RH

To test prediction (3a), for each participant we first calculated average values of $Pr\,[\text{dir}]\,(A|B)$ and $Pr\,[\text{ind}]\,(A|B)$ across

her 12 sets. Then at each level we performed paired t-tests and Wilcoxon signed-rank test on these averages across the 45 participants. In all three cases, $Pr\,[\mathrm{dir}]\,(A|B)$ was reliably smaller than $Pr\,[\mathrm{ind}]\,(A|B)$ (for low, medium, high levels, paired $t(44) = 7.1, 8.4, 3.8$, respectively, $p < .01$; Wilcoxon test also yields $p < .01$ in all three cases). The differences between $Pr\,[\mathrm{dir}]\,(A|B)$ and $Pr\,[\mathrm{ind}]\,(A|B)$ are plotted in Fig. 2, which shows a gap of 0.36, 0.30, and 0.21 at the three levels. The average of $Pr\,[\mathrm{ind}]\,(A|B)$ exceeded the average of $Pr\,[\mathrm{dir}]\,(A|B)$ for 44 of the 45 participants at the low level, 43 at the medium level, and 41 at the high level.

Interpretation of the preceding statistics is clouded by biases that can arise from considering ratios of responses. For example, a given reduction ϵ in $Pr\,[\mathrm{dir}]\,(B)$ can raise $Pr\,[\mathrm{ind}]\,(A|B)$ more than an increase of ϵ in $Pr\,[\mathrm{dir}]\,(B)$ lowers $Pr\,[\mathrm{ind}]\,(A|B)$ (ϵ might represent error in reporting subjective probability). Hence, to document the overestimation of $Pr\,[\mathrm{dir}]\,(A|B)$ by $Pr\,[\mathrm{ind}]\,(A|B)$, we also relied on the following qualitative analyses, not involving means of $Pr\,[\mathrm{dir}]\,(A|B)$ and $Pr\,[\mathrm{ind}]\,(A|B)$. For each participant, we counted the number of trials out of 12 where $Pr\,[\mathrm{ind}]\,(A|B) > Pr\,[\mathrm{dir}]\,(A|B)$. Across the 45 participants, this inequality held for 79.63% of the trials on average (SD = 0.15; reliably greater than 50%, $t(44) = 13.5, p < .01$). Further, for 41 of the 45 participants, $Pr\,[\mathrm{ind}]\,(A|B) > Pr\,[\mathrm{dir}]\,(A|B)$ for at least 7 out of 12 trials ($p \approx 0$ by binomial test).

To test prediction (3b), for each participant and each set we computed the absolute difference between $Pr\,[\mathrm{dir}]\,(A|B)$ and $Pr\,[\mathrm{obj}]\,(A|B)$, and between $Pr\,[\mathrm{ind}]\,(A|B)$ and $Pr\,[\mathrm{obj}]\,(A|B)$. Average absolute differences were then calculated for each participant across her 12 sets. Across the 45 participants, the average absolute difference between $Pr\,[\mathrm{dir}]\,(A|B)$ and $Pr\,[\mathrm{obj}]\,(A|B)$ was 0.10, 0.10, and 0.08 at the three levels, compared to 0.43, 0.37, and 0.20 for $Pr\,[\mathrm{ind}]\,(A|B)$. Thus, direct estimates of objective conditional probability were more accurate than indirect at the low and medium levels (paired t-tests yield $t(44)$=7.2 and 7.9, p<.01; Wilcoxon test yields p<.01). The accuracy of direct estimates was close to that of indirect estimates at the high level ($t(44)$=1.9,p>.05;p>.05 also by Wilcoxon test).

The inaccuracy of $Pr\,[\mathrm{ind}]\,(A|B)$ – namely, $Pr\,[\mathrm{dir}]\,(A \cap B)/Pr\,[\mathrm{dir}]\,(B)$ – as an estimate of $Pr\,[\mathrm{obj}]\,(A|B)$ derives from overestimation of $Pr\,[\mathrm{obj}]\,(A \cap B)$ along with underestimation of $Pr\,[\mathrm{obj}]\,(B)$. Indeed, at low and medium levels, the means of $Pr\,[\mathrm{dir}]\,(A \cap B)$ over the 45 participants were reliably greater than the corresponding means of $Pr\,[\mathrm{obj}]\,(A \cap B)$ ($p < .01$ in both cases). At the high level, $Pr\,[\mathrm{dir}]\,(A \cap B)$ was only slightly greater than $Pr\,[\mathrm{obj}]\,(A \cap B)$ ($p > .05$). These findings are consistent with Bar-Hillel (1973), who argues that conjunctive events have a tendency towards overestimation (whereas disjunctive events have a tendency towards underestimation). $Pr\,[\mathrm{dir}]\,(B)$ was reliably smaller than $Pr\,[\mathrm{obj}]\,(B)$ at medium and high levels ($p < .05, .01$). At the low level, $Pr\,[\mathrm{dir}]\,(B)$ was slightly greater than $Pr\,[\mathrm{obj}]\,(B)$ ($p > .05$).

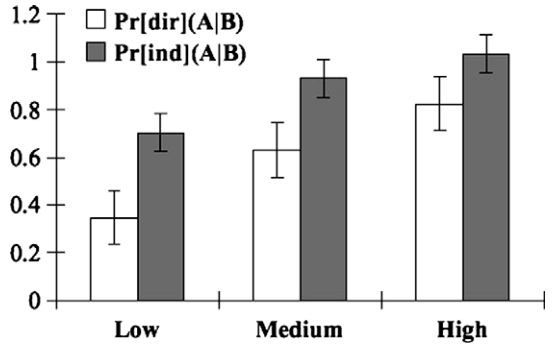

Fig. 2. Comparison of $Pr\,[\mathrm{dir}]\,(A|B)$ to $Pr\,[\mathrm{ind}]\,(A|B)$ in Experiment 1. At all three levels (as specified in Table 2), $Pr\,[\mathrm{ind}]\,(A|B)$ overestimated $Pr\,[\mathrm{dir}]\,(A|B)$. In accordance with RH, $Pr\,[\mathrm{ind}]\,(A|B)$ was computed from Eq. (1).

2.6. Inversion of conditional probability

The data give scant evidence for confusion of $Pr\,(B|A)$ with $Pr\,(A|B)$. For each participant at each level, we calculated the mean absolute difference between $Pr\,[\mathrm{obj}]\,(A|B)$ and $Pr\,[\mathrm{dir}]\,(A|B)$ along with the mean absolute difference between $Pr\,[\mathrm{obj}]\,(B|A)$ and $Pr\,[\mathrm{dir}]\,(A|B)$. These means were based on the eight estimates of $Pr\,[\mathrm{dir}]\,(A|B)$ made at a given level. Inversion of the conditional would result in the absolute difference between $Pr\,[\mathrm{obj}]\,(A|B)$ and $Pr\,[\mathrm{dir}]\,(A|B)$ tending to be equal to or greater than the absolute difference between $Pr\,[\mathrm{obj}]\,(B|A)$ and $Pr\,[\mathrm{dir}]\,(A|B)$. We found, however, that the first difference was smaller than the second at each level. This effect was significant at the low level (paired $t(44) = 2.9, p < .01$) but was just a trend at the medium and high levels (paired $t(44) = 1.4$ in each case, $p > .05$).

2.7. Discussion of Experiment 1

The results of Experiment 1 are inconsistent with RH, the hypothesis that conditional probabilities are mentally calculated from the ratio appearing in Eq. (1). Both quantitative and qualitative analyses reveal that the latter equation overestimates participants' direct judgments of $Pr\,(A|B)$. Qualitative analyses were based on just the direction of misestimation of $Pr\,[\mathrm{dir}]\,(A|B)$ by $Pr\,[\mathrm{ind}]\,(A|B)$ rather than the magnitude (to avoid biases that potentially arise in taking ratios of estimates). Moreover, at the low and medium levels, direct estimates of objective conditional probability were considerably more accurate than estimates based on RH.

Finally, participants showed no sign of conflating $Pr\,(A|B)$ with $Pr\,(B|A)$. Combined with the accuracy of their estimates of $Pr\,[\mathrm{obj}]\,(A|B)$, these results suggest they have a mature conception of conditional probability.

3. Experiment 2

The second experiment was designed to replicate the first, and also to test Hypothesis RH′ [based on Eq. (2)] via its predictions (4). For this purpose, we added the query $Pr\,(\overline{A} \cap B)$ to the three queries figuring in Experiment 1.

3.1. Participants

Forty-five undergraduate students from Princeton University participated in exchange for partial course credit

(35 female, mean age 19.2 years, SD = 1.35). None had participated in Experiment 1.

3.2. Materials and procedure

The stimuli from Experiment 1 were employed again. The procedure was the same except that each set figured in an additional trial that queried $Pr(\overline{A} \cap B)$ as illustrated here.

> PROBABILITY AND FREQUENCY QUERIES FOR $Pr(\overline{A} \cap B)$:
> probability version: What is the probability that a randomly selected shape in the set is square and not red?
> frequency version: What percent of the shapes in the set are square and not red?

Note that these queries have the form $Pr(B \cap \overline{A})$ rather than the equivalent $Pr(\overline{A} \cap B)$. This was done to avoid ambiguity about the scope of the negation. To summarize, in both conditions a given set figured in four trials, one for each of the probabilities $Pr[\text{dir}](B), Pr[\text{dir}](A \cap B), Pr[\text{dir}](\overline{A} \cap B)$ and $Pr[\text{dir}](A|B)$.

3.3. Results

3.3.1. Average responses

Once again, the probability and frequency conditions produced similar numbers; across all participants, the average discrepancy between responses to corresponding queries was only 0.024. Using the same tests as in Experiment 1, we found no significant differences between probability versus frequency responses for any of the four kinds of queries. The two conditions were therefore collapsed as before. For a given participant, we averaged the response to each query – $Pr[\text{dir}](B), Pr[\text{dir}](A \cap B), Pr[\text{dir}](\overline{A} \cap B)$ or $Pr[\text{dir}](A|B)$ – at each level (low, medium, high). Each of these twelve categories of numbers (three levels by four queries) was then averaged across the 45 participants. To compute $Pr[\text{ind}'](A|B)$, for each participant and each set, we divided her estimate of $Pr[\text{dir}](A \cap B)$ by the sum of her estimates of $Pr[\text{dir}](A \cap B)$ and $Pr[\text{dir}](\overline{A} \cap B)$ [just as for $Pr[\text{ind}](A|B)$]. The results are presented in Table 3.

3.4. Replication of Experiment 1

The results of Experiment 1 were replicated in Experiment 2. Regarding prediction (3a) of Hypothesis RH, the average value of $Pr[\text{dir}](A|B)$ across the 45 participants of Experiment 2 was reliably smaller than $Pr[\text{ind}](A|B)$ at all three levels (paired $t(44) = 8.2, 5.8, 3.1$, respectively, $p < .01$; Wilcoxon test also yields $p < .01$ in all three cases). At the three levels, the respective differences between $Pr[\text{ind}](A|B)$ and $Pr[\text{dir}](A|B)$ were 0.32, 0.37, and 0.32. The average of $Pr[\text{ind}](A|B)$ exceeded the average of $Pr[\text{dir}](A|B)$ for 45 of the 45 participants at the low level, 43 at the medium level, and 38 at the high level.

Qualitatively, on average, 81.52% of trials (out of 12) showed $Pr[\text{ind}](A|B) > Pr[\text{dir}](A|B)$ (SD = 0.14; reliably greater than 50%, $t(44) = 15.3, p < .01$). Further, in *all* 45 participants, the latter inequality held in at least 7 out of 12 trials.

Regarding prediction (3b), direct estimates of $Pr[\text{obj}](A|B)$ were closer than indirect estimates at all three levels ($t(44) = 5.3, 4.8, 2.5, p < .01$ for the low and medium levels, $p < .05$ for the high level; Wilcoxon test also yields $p < .01$ for the low and medium levels, $p < .05$ for the high level). Across the 45 participants, the average absolute difference between $Pr[\text{dir}](A|B)$ and $Pr[\text{obj}](A|B)$ was 0.09, 0.12, and 0.07 at the three levels, compared to 0.33, 0.39, and 0.32 for $Pr[\text{ind}](A|B)$. As before, at low and medium levels, the mean of $Pr[\text{dir}](A \cap B)$ was greater than $Pr[\text{obj}](A \cap B)$ ($p < .01$). At medium and high levels, $Pr[\text{dir}](B)$ was smaller than $Pr[\text{obj}](B)$ ($p < .05, .01$, respectively). Thus, the inaccuracy of $Pr[\text{ind}](A|B)$ derives from both its numerator and denominator.

Again, there was little evidence for conflation of $Pr(B|A)$ with $Pr(A|B)$. $Pr[\text{dir}](A|B)$ was significantly closer to $Pr[\text{obj}](A|B)$ than to $Pr[\text{obj}](B|A)$ at the low and high levels (paired $t(44) = 2.8, 3.1$, respectively, $p < .01$), and also closer at the medium level but not significantly (paired $t(44) = 1.1, p > .05$).

3.5. Test of RH′

To test prediction (4a) that $Pr[\text{ind}'](A|B)$ is an unbiased estimate of $Pr[\text{dir}](A|B)$, for each participant we first calculated average values of $Pr[\text{dir}](A|B)$ and $Pr[\text{ind}'](A|B)$ across her 12 sets. Then at each level we performed paired *t*-tests and Wilcoxon signed-rank test on the averages of $Pr[\text{dir}](A|B)$ versus $Pr[\text{ind}'](A|B)$ across the 45 participants. For all three levels, $Pr[\text{dir}](A|B)$ was reliably smaller than $Pr[\text{ind}'](A|B)$ ($t(44) = 6.9, 4.2, 2.5$, respectively, $p < .01$; same for the Wilcoxon test), with gaps of 0.12, 0.08, and 0.03. Although $Pr[\text{ind}'](A|B)$ systematically overestimates $Pr[\text{dir}](A|B)$, its gaps are smaller than for $Pr[\text{ind}](A|B)$. See Fig. 3. The average of $Pr[\text{ind}'](A|B)$ exceeded the average of $Pr[\text{dir}](A|B)$ for 39, 33, and 30 of the 45 participants at low, medium, and high levels, respectively, compared to 45, 43, 38 for $Pr[\text{ind}](A|B)$, as reported above.

Qualitatively, for each participant, we counted the number of trials out of 12 where $Pr[\text{ind}'](A|B) > Pr[\text{dir}](A|B)$. Across 45 participants, this inequality held for 68.56% of

Table 3
Average direct and indirect estimates segregating by level from Experiment 2 (standard deviations in parentheses).

Level	B	$A \cap B$	$\overline{A} \cap B$	$A\|B$	$\text{ind}(A\|B)$	$\text{ind}'(A\|B)$
Low	0.29 (0.05)	0.17 (0.06)	0.23 (0.07)	0.33 (0.09)	0.65 (0.25)	0.45 (0.07)
Medium	0.59 (0.08)	0.51 (0.16)	0.21 (0.05)	0.63 (0.10)	1.00 (0.41)	0.71 (0.07)
High	0.85 (0.05)	0.78 (0.06)	0.14 (0.08)	0.83 (0.06)	1.15 (0.68)	0.86 (0.06)

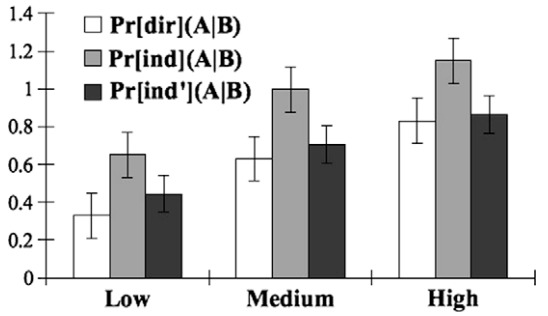

Fig. 3. $Pr\,[\mathrm{dir}]\,(A|B), Pr\,[\mathrm{ind}]\,(A|B)$ and $Pr\,[\mathrm{ind}']\,(A|B)$ in Experiment 2. At all three levels, $Pr\,[\mathrm{ind}']\,(A|B)$ is closer to $Pr\,[\mathrm{dir}]\,(A|B)$ than is $Pr\,[\mathrm{ind}]\,(A|B)$. In accordance with RH and RH′, $Pr\,[\mathrm{ind}]\,(A|B)$ and $Pr\,[\mathrm{ind}']\,(A|B)$ were computed from Eqs. (1) and (2), respectively.

the trials on average (SD = 0.15; reliably greater than 50%, $t(44) = 8.2, p < .01$). For 38 out of 45 participants, $Pr\,[\mathrm{ind}']\,(A|B) > Pr\,[\mathrm{dir}]\,(A|B)$ for at least 7 out of 12 trials ($p < .01$).

To document the greater accuracy of RH′ compared to RH, we counted for each participant the number of trials in which

$$|Pr\,[\mathrm{ind}']\,(A|B) - Pr\,[\mathrm{dir}]\,(A|B)| < |Pr\,[\mathrm{ind}]\,(A|B) - Pr\,[\mathrm{dir}]\,(A|B)| \tag{5}$$

that is, in which $Pr\,[\mathrm{ind}']\,(A|B)$ was closer to $Pr\,[\mathrm{dir}]\,(A|B)$ than $Pr\,[\mathrm{ind}]\,(A|B)$ was. Across 45 participants, (5) held in 69.04% of the trials on average (SD = 0.17; reliably greater than 50%, $t(44) = 7.4, p < .01$). For 34 out of 45 participants, (5) held for at least 7 out of 12 trials ($p < .01$).

RH′ thus appears to be more accurate than RH as a predictor of $Pr\,[\mathrm{dir}]\,(A|B)$. The superiority of RH′ must be due to the denominator $Pr\,(A \cap B) + Pr\,(\overline{A} \cap B)$ in (2) compared to $Pr\,(B)$ in (1) inasmuch as the respective numerators are identical. Indeed, despite the equivalence of $Pr\,(B)$ and $Pr\,(A \cap B) + Pr\,(\overline{A} \cap B)$ in the probability calculus, the mean of $Pr\,[\mathrm{dir}]\,(A \cap B) + Pr\,[\mathrm{dir}]\,(\overline{A} \cap B)$ exceeded the mean of $Pr\,[\mathrm{dir}]\,(B)$ at all three levels ($p < .01$ in each case) (thus, judgment was *subadditive*). The greater value of the denominator in (2) compared to (1) lowers the value of the ratio thereby mitigating the overestimation of $Pr\,[\mathrm{dir}]\,(A|B)$.

To test the prediction (4b) that $Pr\,[\mathrm{dir}]\,(A|B)$ and $Pr\,[\mathrm{ind}']\,(A|B)$ are equally close to $Pr\,[\mathrm{obj}]\,(A|B)$, for each participant and each set we computed the absolute difference between $Pr\,[\mathrm{dir}]\,(A|B)$ and $Pr\,[\mathrm{obj}]\,(A|B)$, and between $Pr\,[\mathrm{ind}']\,(A|B)$ and $Pr\,[\mathrm{obj}]\,(A|B)$. At the low level, $Pr\,[\mathrm{dir}]\,(A|B)$ was significantly more accurate than $Pr\,[\mathrm{ind}']\,(A|B)$ ($t(44) = 2.1, p < .05$; $p < .05$ also by Wilcoxon test) whereas $Pr\,[\mathrm{ind}']\,(A|B)$ was more accurate than $Pr\,[\mathrm{dir}]\,(A|B)$ at the other levels; for the medium level the difference was significant ($t(44) = 2.0, p < .05$; $p < .05$ also by Wilcoxon test) but just a trend at the high level $t(44) = 1.4, p > .05$; $p > .05$ also by Wilcoxon test. Across the 45 participants, the average absolute difference between $Pr\,[\mathrm{dir}]\,(A|B)$ and $Pr\,[\mathrm{obj}]\,(A|B)$ was 0.09, 0.12, and 0.07 at the three levels, compared to 0.13, 0.09, and 0.06 for $Pr\,[\mathrm{ind}']\,(A|B)$.

As a predictor of $Pr\,[\mathrm{obj}]\,(A|B)$, $Pr\,[\mathrm{ind}']\,(A|B)$ was superior to $Pr\,[\mathrm{ind}]\,(A|B)$. In fact, 42, 44, and 35 out of the 45 participants showed smaller average, absolute deviation between $Pr\,[\mathrm{obj}]\,(A|B)$ and $Pr\,[\mathrm{ind}']\,(A|B)$ than between $Pr\,[\mathrm{obj}]\,(A|B)$ and $Pr\,[\mathrm{ind}]\,(A|B)$ at the low, medium, and high levels, respectively.

Finally, we observe that $Pr\,[\mathrm{ind}']\,(A|B)$ predicts $Pr\,[\mathrm{dir}]\,(A|B)$ better than $Pr\,[\mathrm{obj}]\,(A|B)$ predicts $Pr\,[\mathrm{dir}]\,(A|B)$ (a point revisited in the General Discussion). Specifically, we counted for each participant the number of trials in which $Pr\,[\mathrm{ind}']\,(A|B)$ was closer to $Pr\,[\mathrm{dir}]\,(A|B)$ than $Pr\,[\mathrm{obj}]\,(A|B)$ was. Across the 45 participants, $Pr\,[\mathrm{ind}']\,(A|B)$ was closer in 60.19% of the trials on average (SD = 0.22; reliably greater than 50%, $t(44) = 3.1, p < .01$). For 30 out of 45 participants, $Pr\,[\mathrm{ind}']\,(A|B)$ was closer to $Pr\,[\mathrm{dir}]\,(A|B)$ than $Pr\,[\mathrm{obj}]\,(A|B)$ was, for at least 7 out of 12 trials ($p < .05$).

3.6. Discussion of Experiment 2

The results of Experiment 1 were replicated in the present study. $Pr\,[\mathrm{ind}]\,(A|B)$ markedly overestimated $Pr\,[\mathrm{dir}]\,(A|B)$, and also predicted $Pr\,[\mathrm{obj}]\,(A|B)$ less well than $Pr\,[\mathrm{dir}]\,(A|B)$. Also, $Pr\,[\mathrm{dir}]\,(A|B)$ was closer to $Pr\,[\mathrm{obj}]\,(A|B)$ than to $Pr\,[\mathrm{obj}]\,(B|A)$, providing no evidence for systematic conflation of conditionals with their inverse.

The novel finding is the greater accuracy of $Pr\,[\mathrm{ind}']\,(A|B)$ compared to $Pr\,[\mathrm{ind}]\,(A|B)$ at predicting $Pr\,[\mathrm{dir}]\,(A|B)$. Although $Pr\,[\mathrm{ind}']\,(A|B)$ overestimates $Pr\,[\mathrm{dir}]\,(A|B)$, its error is about half that of $Pr\,[\mathrm{ind}]\,(A|B)$ (Fig. 3). This is due to subadditivity: the denominator $Pr\,[\mathrm{dir}]\,(A \cap B) + Pr\,[\mathrm{dir}]\,(\overline{A} \cap B)$ in Eq. (2) exceeds the denominator $Pr\,[\mathrm{dir}]\,(B)$ in Eq. (1).

Moreover, in a significant majority of participants, $Pr\,[\mathrm{ind}']\,(A|B)$ was closer to $Pr\,[\mathrm{dir}]\,(A|B)$ than was $Pr\,[\mathrm{ind}]\,(A|B)$. Note that both $Pr\,[\mathrm{ind}]\,(A|B)$ and $Pr\,[\mathrm{ind}']\,(A|B)$ are computed from ratios, mitigating the impact of ratio biases in this comparison. Finally, unlike $Pr\,[\mathrm{ind}]\,(A|B)$, $Pr\,[\mathrm{ind}']\,(A|B)$ predicts $Pr\,[\mathrm{obj}]\,(A|B)$ about as well as does $Pr\,[\mathrm{dir}]\,(A|B)$. RH′ is thus better supported than RH by our data.

4. Experiment 3

The sample space in the previous two experiments is transparent, and all probabilities were grounded in frequencies. The results thus test RH and RH′ when probability is *extensional*. The third experiment was designed to evaluate RH and RH′ in an *intensional* setting which involved probabilities of non-repeatable events (conceived as statements).[3]

In this setting it is not feasible to query a given participant about $Pr\,(A|B), Pr\,(A \wedge B), Pr\,(\neg A \wedge B)$, and $Pr\,(B)$ for the same events A, B, without her knowing it (unlike the two preceding experiments, where this *was* possible). We therefore relied here on a between-subjects design.

4.1. Participants

Five hundred and fifty-seven people from the general public took part in the experiment (289 female, mean

[3] Pedantry: because A, B are here interpreted as statements rather than events, we write $Pr\,(A \wedge B)$ in place of the set-theoretic $Pr\,(A \cap B)$.

age 32.54 years, SD = 12.42). They were recruited anonymously through the internet via *Amazon Mechanical Turk*.

4.2. Materials

Twelve pairs of future events were constructed (see the Appendix), for example:

$$\begin{aligned} A &= \text{Humans walk on Mars by 2050.} \\ B &= \text{NASA merges with the European Space Agency by 2030.} \end{aligned} \quad (6)$$

For a given pair, four statements of the forms $B, A \wedge B, \neg A \wedge B$, and $A|B$ were generated, for example:

Sample statements for one set:

B	NASA merges with the European Space Agency by 2030.
$A \wedge B$	Humans walk on Mars by 2050 and NASA merges with the European Space Agency by 2030.
$\neg A \wedge B$	Humans do not walk on Mars by 2050 and NASA merges with the European Space Agency by 2030.
$A\|B$	Humans walk on Mars by 2050, assuming that NASA merges with the European Space Agency by 2030.

4.3. Procedure

For each participant, a set of 12 statements was constructed by choosing one statement from each of the 12 blocks of four (illustrated above). Each participant was invited to supply her personal probability for each of her 12 statements. Thus, each participant evaluated just one of $Pr(A|B), Pr(A \wedge B), Pr(\neg A \wedge B)$, and $Pr(B)$ for any given pair A, B of events.

4.4. Results

For given events A, B, group averages were used to define $Pr[\text{dir}](B), Pr[\text{dir}](A \wedge B), Pr[\text{dir}](\neg A \wedge B)$, and $Pr[\text{dir}](A|B)$. $Pr[\text{ind}](A|B)$ and $Pr[\text{ind}'](A|B)$ were then computed via Eqs. (1) and (2), respectively (there are no objective probabilities in the present setting). Results are shown in Table 4. The table provides striking illustration of subadditivity. For all 12 items, $Pr[\text{dir}](A \wedge B) + Pr[\text{dir}](\neg A \wedge B) > Pr[\text{dir}](B)$.

4.5. Test of RH

To test whether $Pr[\text{ind}](A|B)$ is an unbiased estimate of $Pr[\text{dir}](A|B)$, we performed an independent t-test on the average values of $Pr[\text{dir}](A|B)$ versus $Pr[\text{ind}](A|B)$, collapsing over the 12 sets. $Pr[\text{dir}](A|B)$ was significantly smaller than $Pr[\text{ind}](A|B)$ ($t(22) = -8.27, p < .01$). Across the 12 sets, the gap between $Pr[\text{dir}](A|B)$ and $Pr[\text{ind}](A|B)$ was an enormous 0.46 on average. Moreover, the average of $Pr[\text{ind}](A|B)$ exceeded the average of $Pr[\text{dir}](A|B)$ for all 12 sets.

4.6. Test of RH′

Test of $Pr[\text{ind}](A|B)$ as an unbiased estimate of $Pr[\text{dir}](A|B)$ proceeded in the same way. We performed an independent t-test on the average values of $Pr[\text{dir}](A|B)$ versus $Pr[\text{ind}'](A|B)$, collapsing across the 12 sets. This time, no significant difference was found. Indeed, the gap between $Pr[\text{dir}](A|B)$ and $Pr[\text{ind}'](A|B)$ was only 0.05, and the average of $Pr[\text{ind}'](A|B)$ exceeded the average of $Pr[\text{dir}](A|B)$ for just seven sets.

4.7. Discussion of Experiment 3

The results replicate Experiments 1 and 2 insofar as $Pr[\text{ind}](A|B)$ grossly overestimates $Pr[\text{dir}](A|B)$. The generality of this finding is underlined by its presence in both the extensional setting of the first two experiments and the intensional setting here. A different replication concerns $Pr[\text{ind}](A|B)$ versus $Pr[\text{ind}'](A|B)$ as predictors of $Pr[\text{dir}](A|B)$. In Experiment 2, $Pr[\text{ind}'](A|B)$ was seen to be more accurate than $Pr[\text{ind}](A|B)$ in this regard. The present results confirm this finding in the intensional context. Indeed, Table 4 reveals Eq. (2) to be remarkably accurate in predicting direct estimates of conditional probability.

5. General discussion

Our findings suggest that judgments of conditional probability do not arise from mental division of the kind envisioned in the standard definition. For, the ratio $Pr(A \cap B)/Pr(B)$ seen in Eq. (1) systematically overestimates such judgments in all three of our experiments. Compared to direct estimates, the ratio is also further from the objective conditional probabilities inherent in the stimuli of the first two experiments, providing another perspective on the limitations of (1) as a psychological theory.

Why might (1) be inaccurate? One possibility is faulty division of the independent estimates of $Pr(A \cap B)$ and $Pr(B)$. Another possibility is that the latter estimates are not independent when performed in the service of calculating $Pr(A|B)$, hence they differ from estimates that result from soliciting each individually. An alternative conjecture is that people rely more on (2) than (1), while their judgments fail to respect the equivalence between $Pr(B)$ and $Pr(A \cap B) + Pr(\overline{A} \cap B)$. This idea is consistent with the greater accuracy of RH′ to RH in Experiment 2, and the impressive accuracy of RH′ (but not RH) in Experiment 3. In turn, the superiority of RH′ to RH is due to the higher estimate of $Pr(B)$ when it is decomposed as $Pr(A \cap B) + Pr(\overline{A} \cap B)$ (subadditivity). As discussed in the Introduction, such decomposition often (but not invariably) increases estimates of event probability (Sloman et al., 2004; Tversky & Koehler, 1994).

In the extensional setting of Experiments 1 and 2, it is easy to envision theories of conditional probability that are alternative to the ratio accounts (1) and (2). Asked about $Pr(\text{red}|\text{square})$, for example, one might attempt to bring to mind just the squares then estimate the propor-

Table 4
Average direct and indirect estimates from Experiment 3 (standard deviations in parentheses).

Set	B	$A \wedge B$	$\neg A \wedge B$	$A\|B$	$\mathrm{ind}(A\|B)$	$\mathrm{ind}'(A\|B)$
1	0.45 (0.29)	0.47 (0.29)	0.43 (0.26)	0.50 (0.30)	1.03	0.52
2	0.42 (0.32)	0.33 (0.29)	0.46 (0.31)	0.42 (0.33)	0.79	0.42
3	0.45 (0.30)	0.46 (0.29)	0.45 (0.27)	0.56 (0.30)	1.03	0.51
4	0.22 (0.20)	0.19 (0.16)	0.32 (0.30)	0.19 (0.17)	0.86	0.37
5	0.68 (0.29)	0.48 (0.30)	0.66 (0.27)	0.44 (0.29)	0.71	0.42
6	0.52 (0.26)	0.60 (0.26)	0.38 (0.25)	0.74 (0.23)	1.15	0.61
7	0.36 (0.27)	0.37 (0.30)	0.37 (0.28)	0.38 (0.31)	1.03	0.50
8	0.72 (0.26)	0.66 (0.24)	0.53 (0.28)	0.55 (0.27)	0.91	0.55
9	0.49 (0.30)	0.40 (0.29)	0.44 (0.29)	0.41 (0.29)	0.80	0.47
10	0.50 (0.27)	0.50 (0.28)	0.42 (0.26)	0.48 (0.29)	1.00	0.54
11	0.34 (0.27)	0.43 (0.30)	0.37 (0.30)	0.41 (0.29)	1.23	0.53
12	0.26 (0.21)	0.22 (0.20)	0.26 (0.18)	0.26 (0.22)	0.84	0.46
Overall	0.46 (0.32)	0.43 (0.31)	0.42 (0.30)	0.45 (0.32)	0.91 (0.16)	0.50 (0.07)

Note: $N = 138$ for sets 1–3. $N = 171$ for sets 4–7. $N = 147$ for sets 7–9. $N = 101$ for sets 10–12. Observe how close the values in the last column are to $A|B$.

tion of reds in this set. When the underlying partition of events is less evident than here it may be challenging to identify the relevant symmetries, opening the door to misconceptions and biases (Fox & Levav, 2004). Event-counting may nonetheless be central to many extensional settings, in which probability can be defined from frequency.

Still, the hypothesis of event-counting needs further specification before it can be used to predict estimates of conditional probability. Pursuing our example, how is the proportion of reds in the set of squares to be determined? Is the idea to compare the number of red squares to the number of squares? This amounts to RH, which did not fare well in our experiments. More generally, which set of squares is held in mind, and which subset of reds? A simple answer is that *the actual set* of presented squares is mentally represented, along with the actual subset of reds. This version of the theory implies that the estimated conditional probability of red given square equals the objective conditional probability. But we saw at the end of the results section for Experiment 2 that $Pr\,[\mathrm{ind}'](A|B)$ predicts $Pr\,[\mathrm{dir}](A|B)$ reliably better than $Pr\,[\mathrm{obj}](A|B)$ predicts $Pr\,[\mathrm{dir}](A|B)$, hence that RH′ is superior to this simple event-counting model as a predictor of estimated conditional probability.

Hypotheses based on event-counting face further challenges in the *intensional case*, involving non-repeatable events like (6) in Experiment 3, repeated here.

$A =$ Humans walk on Mars by 2050.

$B =$ NASA merges with the European Space Agency by 2030.

To apply event-counting, it seems necessary to posit a set of *possible worlds* of equal positive probability, obtaining $Pr\,(A|B)$ by counting the worlds that satisfy A among those that satisfy B. Such a scheme might be plausible when boolean connectives are in play (Johnson-Laird, 2006) but for A, B above there seems to be no limit on how many possibilities can be imagined in which B is true.

Instead of counting worlds, perhaps we mentally represent the broad contours of a few salient possibilities that satisfy B, estimate the probability of A in each of them, then average. For example, one scenario is that the merger foretold by B arises from competition with Asian space programs, another that it ensues from budgetary constraints in the US (often just a single scenario might come to mind). A limitation of this theory, however, is that it begs the question of how the probability of A is evaluated in a given mental scenario that satisfies B. Indeed, one scenario is "the actual state of affairs except that B holds". Estimating $Pr\,(A)$ therein is the same as estimating $Pr\,(A|B)$, bringing us full circle.[4]

Another idea is that $Pr\,(A|B)$ derives from weighing the chance that a particular conditional statement is true, perhaps:

Were NASA to merge with the European Space Agency by 2030, humans would walk on Mars by 2050. (7)

This line of thought encounters two objections. First, it is notoriously difficult to elucidate the meanings of sentences like (7), in particular, to provide a tractable account of their truth and falsity (Harper, Stalnaker, & Pearce, 1981; Sanford, 2003). In the absence of such an account it is unclear what probability is at issue when considering the chance of (7).[5] Second, conceiving $Pr\,(A|B)$ in this way is a reduction of *binary* conditional probability to *unary* absolute probability. The reduction is achieved by transforming the pair A, B into a single sentence $f(A, B)$ via some grammatical operation f, e.g., conversion of A and B to subjunctive/conditional tense as in (7). It is well known, however, that there is no such transformation with $Pr\,(A|B) = Pr\,(f(A, B))$ (Bradley, 1999; Lewis et al., 1976).

The foregoing fact is compatible with estimating the probability of a sentence like (7) to be $Pr\,(A|B)$. Indeed, in

[4] The scheme just outlined receives more refined expression in David Lewis' imaging principle and variants thereof Lewis et al. (1976). Imaging, however, is a complicated construction, relying on a similarity metric among scenarios.

[5] Some authors distinguish between the probability of a conditional C and the probability that C is true but we find this distinction difficult to interpret. For discussion, see Lycan (2001, chap. 4).

extensional settings involving indicative conditionals (if p then q) people often make this choice (Evans, Handley, & Over (2003), consistent with the theory of Adams (1975)). It is the converse hypothesis that seems problematic, according to which people evaluate the conditional probability of A given B by first constructing a sentence like (7) then evaluating it absolutely. To the extent that judgment satisfies the axioms of probability, conditional probabilities cannot in general be computed this way.

A different approach is to consider $Pr(A|B)$ undefined unless B is true (or believed so), in which case the probability of $Pr(A|B)$ is $Pr(A)$. This idea makes sense of the kind of *betting contract* often associated with conditional probabilities (Hacking, 2001). But it seems unacceptable as a psychological theory inasmuch as people are typically willing to estimate $Pr(A|B)$ without first accepting B.

None of the difficulties discussed above afflict the hypothesis that $Pr(A|B)$ is mentally computed from Eq. (2). The ratio $Pr(A \wedge B)/[Pr(A \wedge B) + Pr(\neg A \wedge B)]$ is a binary function involving no conditionals, and is defined provided only that $Pr(B) > 0$. The latter proviso seems more palatable in the intensional case compared to the counterintuitive results it produces extensionally (Hájek, 2003). Unfortunately, RH′ does not illuminate how $Pr(A \wedge B)$ or $Pr(\neg A \wedge B)$ are calculated, which appears to be just as mysterious in the intensional framework as the calculation of $Pr(A|B)$. Both require determining the compatibility of A and B.

Finally, we underline the preliminary nature of the present study. A wide range of stimuli will be necessary to reach firm conclusions about the mental representation of conditional probability. Indeed, different representations could be evoked by different kinds of events. Results might also depend on the protocol used to elicit probabilities. Here we relied on direct estimation (using both frequency and probability formulations). Our findings (in particular, the relative accuracy of RH′) might be different if probabilities were measured via betting rates. Explicit and quantitatively exact alternatives to RH′ would also be valuable. Although surprisingly accurate in the present study (especially in the third experiment), the kind of decomposition and division embodied by Eq. (2) may not represent the mental steps involved in constructing conditional probability.

Acknowledgement

Thanks to Steven Sloman and David Over for helpful comments on an earlier version of this manuscript. Three anonymous referees also provided valuable input. Osherson acknowledges support from the Henry Luce Foundation. Contact information: jiayingz/akshah/osherson@princeton.edu.

Appendix. Twelve pairs of future events used in Experiment 3.

Set 1
A = Humans walk on Mars by 2050.
B = NASA merges with the European Space Agency by 2030.

Set 2
A = John McCain campaigns for Sarah Palin in 2012.
B = Sarah Palin runs for President in 2012.

Set 3
A = Humans colonize an extraterrestrial body by 2016.
B = Extraterrestrial life is discovered by 2016.

Set 4
A = A fully featured laptop retails for $100 or less in the US by 2015.
B = Dell declares bankruptcy by 2015.

Set 5
A = Dow Jones falls below 6,000 sometime this year.
B = The US government injects more than $25 billion into Big 3 automakers by the end of this year.

Set 6
A = The average life expectancy in the world increases by 10% by 2020.
B = A cure for AIDS is discovered by 2020.

Set 7
A = Ford releases a new model of hydrogen vehicle by the end of 2010.
B = The national average for Regular Gasoline settles above $2.00 per gallon by the end of 2010.

Set 8
A = A vaccine to avian influenza is discovered by 2012.
B = Avian influenza reaches the US by 2012.

Set 9
A = The number of casinos in Las Vegas increases by 20% by 2015.
B = The use of marijuana becomes legalized in the US by 2015.

Set 10
A = Private investors obtain lunar property rights by 2012.
B = Water is discovered on the moon by 2012.

Set 11
A = The US dollar reaches parity with the Euro (1 dollar=1 euro) by 2015.
B = The US unemployment rate reaches 15% in 2015.

Set 12
A = President Obama gets re-elected in 2012.
B = President Obama's approval rating is above 50% at the end of 2012.

References

Adams, E. (1975). *The logic of conditionals*. Dordrecht: Reidel.
Bar-Hillel, M. (1973). On the subjective probability of compound events. *Organizational Behavior and Human Performance, 9*, 396–406.
Bernardo, J., & Smith, A. (1994). *Bayesian theory*. New York: John Wiley and Sons.
Bradley, R. (1999). More triviality. *Journal of Philosophical Logic, 28*(2), 129–139.
Brenner, L., & Rottenstreich, Y. (1999). Focus, repacking and the judgment of grouped hypotheses. *Journal of Behavioral Decision Making, 12*, 141–148.
Brenner, L. A., & Koehler, D. J. (1999). Subjective probability of disjunctive hypotheses: Local-weight models for decomposition of evidential support. *Cognitive Psychology, 38*, 16–47.
Cohen, J., Dearnaley, E., & Hansel, C. (1956). The addition of subjective probabilities: The summation of estimates of success and failure. *Acta Psychologica, 12*, 371–380.

Dawes, R., Mirels, H. L., Gold, E., & Donahue, E. (1993). Equating inverse probabilities in implicit personality judgments. *Psychological Science, 4*(6), 396–400.

Eddy, D. M. (1982). Probabilistic reasoning in clinical mediciine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Evans, J. St. B. T., Handley, S. J., & Over, D. E. (2003). Conditionals and conditional probability. *Journal of Experimental Psychology, 29*(2), 321–335.

Feeney, A., & Heit, E. (Eds.). (2007). *Inductive reasoning: Experimental, developmental, and computational approaches*. Cambridge UK: Cambridge University Press.

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

Fischhoff, B., & Bruine de Bruin, W. (1999). Fifty-fifty=50? *Journal of Behavioral Decision Making, 12*, 149–167.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance, 4*, 330–344.

Fox, C. R., & Levav, J. (2004). Partitioneditcount: Naive extensional reasoning in judgment of conditional probability. *Journal of Experimental Psychology: General, 133*(4), 626–642.

Hacking, I. (2001). *An introduction to probability and inductive logic*. Cambridge UK: Cambridge University Press.

Hájek, A. (2003). What conditional probability could not be. *Synthese, 137*, 273–323.

Harper, W. L., Stalnaker, R., & Pearce, G. (Eds.). (1981). *Ifs*. Boston, MA: D. Reidel.

Idson, L. C., Krantz, D. H., Osherson, D., & Bonini, N. (2001). The relation between probability and evidence judgment: An extension of support theory. *Journal of Risk and Uncertainty, 22*(3), 227–249.

Jeffrey, R. C. (2004). *Subjective probability: The real thing*. Cambridge: Cambridge University Press.

Johnson-Laird, P. N. (2006). *How we reason*. Oxford University Press.

Koehler, J. (1996). The base rate fallacy reconsidered: Descriptive normative and methodological challenges. *Behavioral and Brain Sciences, 19*, 1–53.

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philosophical Review, 85*, 297–315.

Lycan, W. G. (2001). *Real conditionals*. Oxford, UK: Oxford University Press.

Macchi, L., Osherson, D., & Krantz, D. H. (1999). Superadditive probability judgment. *Psychological Review, 106*(1), 210–214.

Mellers, B. A., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science, 12*, 269–275.

Murphy, G. L., & Ross, B. H. (2005). The two faces of typicality in category-based induction. *Cognition, 95*, 175–200.

Popper, K. (1959). *The logic of scientific discovery*. New York: Basic Books.

Ross, B. H., & Murphy, G. L. (1996). Category-based predictions: Influence of uncertainty and feature associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(3), 736–753.

Ross, B. H., & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology, 38*, 495–553.

Ross, S. (1988). *A first course in probability* (3rd ed.). New York City: Macmillan.

Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review, 104*, 406–415.

Russo, J. E., & Kolzow, K. J. (1994). Where is the fault in fault trees? *Journal of Experimental Psychology: Human Perception and Performance, 20*, 17–32.

Sanford, D. H. (2003). *If P, then Q: Conditionals and the foundations of reasoning* (2nd ed.). London: Routledge.

Sloman, S. A., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., & Fox, C. R. (2004). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(3), 573–582.

Tentori, K., Bonini, N., & Osherson, D. (2004). The conjunction fallacy: A misunderstanding about conjunction? *Cognitive Science, 28*(3), 467–477.

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). New York, NY: Cambridge University Press.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*(4), 547–567.

Weber, M., & Osherson, D. (in press). Similarity and induction. *European Review of Philosophy and Psychology*.