Cognition 120 (2011) 248–267

Contents lists available at ScienceDirect

Cognition

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

Reducing cognitive biases in probabilistic reasoning by the use of logarithm formats

Peter Juslin *, Håkan Nilsson, Anders Winman, Marcus Lindskog

Department of Psychology, Uppsala University, Box 1225, SE-751 42 Uppsala, Sweden

ARTICLE INFO

Article history:
Received 7 February 2011
Revised 5 May 2011
Accepted 6 May 2011
Available online 2 June 2011

Keywords:
Probability judgment
Base-rate neglect
Linear models

ABSTRACT

Research on probability judgment has traditionally emphasized that people are susceptible to biases because they rely on "variable substitution": the assessment of normative variables is replaced by assessment of heuristic, subjective variables. A recent proposal is that many of these biases may rather derive from constraints on cognitive integration, where the capacity-limited and sequential nature of controlled judgment promotes linear additive integration, in contrast to many integration rules of probability theory (Juslin, Nilsson, & Winman, 2009). A key implication by this theory is that it should be possible to improve peoples' probabilistic reasoning by changing probability problems into logarithm formats that require additive rather than multiplicative integration. Three experiments demonstrate that recasting tasks in a way that allows people to arrive at the answers by additive integration decreases cognitive biases, and while people can rapidly learn to produce the correct answers in an additive formats, they have great difficulty doing so with a multiplicative format.

1. Introduction

> "*Human judgment is a cognitive activity of last resort*" (K. R. Hammond)

For more than 30 years, research on human probability judgment in cognitive psychology has been dominated by two related claims: First, that people are often poor at assessing probabilities and reasoning with probabilities, and; second, that this is explained by people substituting the hard facts of probability (e.g., frequencies and set sizes) with subjective, "intensional", variables that are conveniently available (e.g., similarity, fluency) by a process referred to as "natural assessment" (Kahneman & Frederick, 2002; see Gilovich, Griffin, and Kahneman (2002) and Kahneman, Slovic, and Tversky (1982) for reviews). Because people are responding to these easily accessible, but normatively inappropriate, subjective variables the judgments violate probability theory, producing a number of judgment biases.

While not denying that people respond to similarity and fluency when making probability judgments, we have recently proposed that the biases may derive not primarily from use of these heuristics per se, but from the pervasive inclination for use of linear additive integration of information (see Juslin et al., 2009; Nilsson, Winman, Juslin, & Hansson, 2009). This argument, too, comes in two related claims: First, because people are capacity-limited, sequential information processors, intuitive judgment tends to implement a linear additive integration of information (Anderson, 1981, 1996; Hogarth & Einhorn, 1992; Juslin, Karlsson, & Olsson, 2008; Lopes, 1985, 1987; Shanteau, 1970, 1972, 1975). This also serves to link research on probabilistic reasoning to the extensive research on linear models in multiple cue judgment (e.g., Hammond & Stewart, 2001; Karelaia & Hogarth, 2008).

* Corresponding author. Address: Department of Psychology, Uppsala University, Box 1225, SE-751 42 Uppsala, Sweden.
E-mail address: peter.juslin@psyk.uu.se (P. Juslin).

0010-0277/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2011.05.004

The second claim is that, to the extent that people base their judgments on noisy input (e.g., small samples), linear additive integration often yields as accurate judgments as reliance on probability theory, which may explain why the mind has evolved with little appreciation for many of the normative coherence rules of probability theory (Juslin et al., 2009). On this view, violations of probability theory derive from processes well adapted both to the cognitive constraints of the human mind and the requirements of noisy real-life environments.

The purpose of this article is to test one key implication of this theory: it should be possible to improve peoples' probabilistic reasoning by changing probability problems into logarithm formats that require additive rather than multiplicative integration. In this context, we benefit from the fact that multiplication in one metric becomes addition, if this metric is represented in terms of its logarithm. Recasting tasks that involve multiplication, and which are associated with classic cognitive biases like *base-rate neglect* (Barbey and Sloman (2007), Kahneman and Tversky (1973), Koehler (1996), and Tversky and Kahneman (1982) for reviews) and the *conjunction fallacy* (Tversky and Kahneman (1983), Costello (2009), Nilsson et al. (2009), and Wedell and Moro (2008) for recent reviews), to allow people to arrive at the correct answer by linear additive integration, should thus immediately decrease the nominal rate of these biases. Moreover, if the linear additive integration, at least in part, arises from cognitive constraints (Juslin et al., 2009) people should easily adapt to produce the correct answer with a linear additive format, but have great difficulty with doing so with a multiplicative format.

We first describe the judgment tasks. Thereafter, to articulate why they are not easily digested by with the human mind, a theoretical framework for human judgment is outlined. In three experiments, we then use these insights to control the rate of cognitive bias.

2. Multiplication with probability theory

Consider a Bayesian inference task like the medical diagnosis problem:

> The probability that a person randomly selected from the population of all Swedes has the disease is 2%. The probability of receiving a positive test result given that one has the disease is 96%. The probability of receiving a positive test result if one does not have the disease is 8%. What is the probability that a randomly selected person that has received a positive test result has the disease?

The posterior probability $p(D|P)$ of Disease given a Positive test is given by Bayes' theorem,

$$p(D|P) = \frac{p(D) \times p(P|D)}{p(D) \times p(P|D) + p(\bar{D}) \times p(P|\bar{D})}, \quad (1)$$

which can be written in its ratio form,

$$\frac{p(D|P)}{p(\bar{D}|P)} = \frac{p(D)}{p(\bar{D})} \times \frac{p(P|D)}{p(P|\bar{D})}, \quad (2)$$

to emphasize the need for a multiplication of base-rate (prior odds ratio) and case evidence (likelihood ratio).[1] $p(D)$ is the base-rate of disease, $p(\bar{D})$ the base-rate of no disease, $p(P|D)$ the probability of a positive test if one has the disease (hit-rate), and $p(P|\bar{D})$ the probability of a positive test if one does not have the disease (false-alarm rate). Typically, the assessed probability is much higher than implied by Bayes' theorem (here .20), often closer to the hit-rate .96, commonly interpreted as a captivation by the hit-rate at the neglect of the low base-rate (.02) (see Eddy, 1982; Gigerenzer & Hoffrage, 1995; Koehler, 1996).

For illustration of another classic cognitive bias that is strongly related to the requirement for multiplicative integration, consider the conjunction fallacy. According to probability theory, the probability of a conjunction, $p(A\&B)$, can never exceed the probability of either of the constituent probabilities, $p(A)$ or $p(B)$, as is most easily seen in the case where the two constituent events A and B are independent, where

$$p(A\&B) = p(A) \times p(B). \quad (3)$$

Because probabilities fall between 0 and 1, clearly $p(A\&B)$ can never exceed $p(A)$ or $p(B)$. The integration rule in Eq. (3) changes if it is assumed that events A and B are dependent, but it is still multiplicative and it remains true that $p(A\&B)$ can never exceed $p(A)$ or $p(B)$. People, however, frequently violate this conjunction rule, as in the classical 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 antinuclear demonstrations. What is the probability of each of the following?
> (A) Linda is active in the feminist movement.
> (B) Linda is a bank teller.
> (C) Linda is a bank teller and is active in the feminist movement.

The conjunction fallacy is the result that the conjunction, *bank teller and feminist*, is generally assessed as more likely than its component *bank teller* (Tversky & Kahneman, 1983).

The original explanations of base-rate neglect and the conjunction fallacy often emphasized the *representativeness heuristic* (Kahneman et al., 1982). People were claimed to assess probabilities by reliance on the similarity or representativeness of the case evidence to the prototypical members of the categories. For example, because Linda appears more representative of feminist bank tellers than of bank tellers, the conjunction is assessed as more likely than the conjunct. The explanations in terms of representativeness have, however, been undermined by more recent studies showing that base-rate neglect and the conjunction fallacy are just as common in tasks where the representativeness heuristic does not apply (e.g., the above medical

[1] Denoting the left-hand posterior odds ratio in Eq. (2) R, the posterior probability $p(D|P)$ is obtained from Eq. (2) by $p(D|P) = R/(R+1)$.

diagnosis problem[2] or Gavanski & Roskos-Ewoldsen's, 1991, and Nilsson, 2008 for, "mixed versions" of the Linda problem). In the medical diagnosis problem, it has instead been proposed that people rely on a "Fisherian algorithm" (Gigerenzer & Hoffrage, 1995), committing "the inverse fallacy" (Villejoubert & Mandel, 2002) essentially reporting the likelihood probability ($p(P|D)$) rather than the posterior probability ($p(D|P)$).

Both base-rate neglect and the conjunction fallacy can be reduced by assessment formats that make it clearer what information that is needed to solve the task. The most common, but not the only, way to do so is to frame the task in terms of natural frequencies (Gigerenzer & Hoffrage, 1995). According to the natural frequency hypothesis, the world reveals itself in the form of natural frequencies and, as a result, humans are adapted to work with natural frequencies (see also Hertwig & Gigerenzer, 1999). When tasks involve single event probabilities, as in the medical diagnosis task or in the Linda task, people fail because they have no tools for processing such information. While this hypothesis can explain why the natural frequency format can be used to reduce many cognitive illusions, it cannot explain why also other formats can reduce the cognitive illusions (Barbey & Sloman, 2007).

The dual system nested set hypothesis of Barbey and Sloman (2007) and the dual system denominator neglect hypothesis of Reyna and colleagues (e.g., Reyna & Mills, 2007; Wolfe & Reyna, 2010) shares several assumptions. For effective reasoning, identification of the sets of events that are relevant to the task is crucial. If these cannot be properly identified, judgments will be based on the wrong information. In the medical diagnosis and the Linda problems, identification of relevant sets and set-relations is difficult for at least two reasons: information is provided as single event probabilities and the relevant sets are nested. Natural frequency is one format that serves to simplify identification of the relevant sets.

Both dual system hypotheses attribute cognitive biases to the intuitive system. The difference is that they provide different explanations of why the cognitive biases are reduced when the relevant sets are defined. Barbey and Sloman (2007) assume that the controlled analytical system works with simple elementary set operations. If essential sets can be identified, the analytical system will get involved and judgments will be (at least fairly) correct. If sets cannot be identified, the intuitive system gets involved and it is the tools of this system that causes illusions. Reyna et al. assume that it is the intuitive system that performs operations on the identified sets. If sets are not identified properly, operations will be performed on the wrong information and this is what causes cognitive illusions. The nested set hypothesis and the denominator neglect hypothesis will not be directly tested in the experiment below. We will, however, return to them in Section 7.

3. Three cognitive layers of human judgment

In order to articulate the relationship between the computational demands when reasoning with probability and the inherent capabilities of the human mind we refer to the theoretical framework in Fig. 1. As illustrated in Fig. 1, the task is conceptualized as involving consideration of one or several cues to make a judgment of some property in the environment. The framework describes three basic cognitive processes that can be used for judgment, roughly corresponding to reasoning, intuitive judgment, and memory.

While analytic judgment processes, corresponding to reasoning, fall neatly into the category of "analytical thinking" and exemplar memory falls into (but does not exhaust) the category of "intuitive thinking", as they occur in current dual systems theories (see Darlow & Sloman, 2010; Evans, 2008; Hammond, 1996), the intuitive judgments in Fig. 1 occupy a middle ground, where the cues are attended and constrained by working memory, but the cue integration is intuitive, arising from architectural constraints of the mind.[3] In the following, we discuss these three processes, concentrating on intuitive judgments of probability.

3.1. Analytic judgment

In some tasks, the judgments can be made by retrieving declarative knowledge of arithmetic facts and analytical principles, from which deductions can be made, as far as it is allowed by the working memory capacity. For example, in a task that requires assessment of a conjunctive probability $p(A\&B)$, where the independent constituent probabilities are $p(A) = .9$ and $p(B) = .1$, the task can be solved by retrieval of (i) the analytical rule dictating that for independent events the probability of the conjunction is the product of the constituents (Eq. (3)) and (ii) the declarative fact that ".9·.1 = .09".Most of these facts are likely to be culturally transmitted by education, both arithmetic facts (e.g., about the "multiplication table") and integration rules (e.g., multiplication for independent events, Bayes' theorem).

Without computers or paper and pencil, one important source of error is the severe limits of working memory (Cowan, 2001; Jonides et al., 2008). While people may be able to mentally add or multiply two digits according to analytic rules by means of controlled thought ("number crunch", as exemplified in the last paragraph), previous

[2] It could be argued that also in the medical diagnosis task, the evidence (a positive test) is more representative of the prototypical person with a disease than of a prototypical person without disease. As noted previously (Gigerenzer & Murray, 1987) this makes the application of representativeness almost circular (i.e., explaining that people only respond to the hit-rate by assuming that they respond to the hit-rate). It appears redundant as compared to characterize such a behavior in terms of a Fisherian algorithm (Gigerenzer & Hoffrage, 1995).

[3] Although there are similarities to the dual-systems theories reviewed in Evans (2008) and Darlow and Sloman (2010), we would also like to emphasize that our framework is not strongly committed to the notion of multiple independent "systems", as typically discussed. What we intend is that people can engage in at least three different kinds of cognitive processes to address a judgment task, and whether these different processes are best conceived of as aspects of one system or as multiple systems is open for scrutiny in future research.

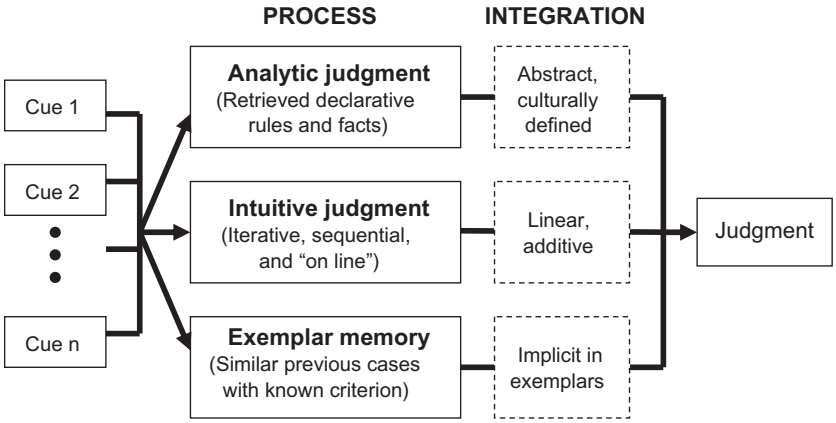

Fig. 1. Framework for human judgment suggesting three alternative cognitive processes by which one or several cues can be transformed into a judgment.

studies clearly suggest that already the explicit numeric reasoning implied by Bayes' theorem in Eq. (1) is beyond the ability of most people (e.g., Eddy, 1982; Gigerenzer & Hoffrage, 1995). If declarative facts are unknown, the task fails to elicit their retrieval, or if the task is too complex to be mastered by elaboration of declarative facts in working memory, the claim is that the judge has to resort to other cognitive processes, like controlled intuitive judgment or exemplar memory.

3.2. Controlled intuitive judgment

As illustrated in the middle-route in Fig. 1, one judgment process involves the controlled, deliberate but intuitive consideration of cues; *controlled* in the sense that it is constrained by working memory, sequential, and that each cue is attended; *intuitive* in the sense that the cue integration is not guided by an explicit analytical rule, like Bayes' theorem. Numerous studies suggest that intuitive judgment is biased towards linear additive cue integration, often as a weighted average (Anderson, 1996; Hogarth & Einhorn, 1992; Juslin et al., 2008; Lopes, 1985, 1987; Roussel, Fayol, & Barrouillet, 2002; Shanteau, 1970, 1972, 1975). Research on multiple-cue judgment, the classic paradigm to study this process, also shows that judgment is often a linear additive function of the cues (Brehmer, 1994; Brunswik, 1952; Cooksey, 1996; Hammond & Stewart, 2001; Karelaia & Hogarth, 2008).

A plausible cognitive explanation for the inclination for weighted and additive integration (Juslin et al., 2008) is that controlled judgment involves an iterative sequential adjustment (e.g., Denrell, 2005; Hogarth & Einhorn, 1992; Lopes, 1985; Shanteau, 1970). At each moment, a new piece of information (cue) is attended, which instills an adjustment of the current estimate of the criterion into a new estimate; then the next cue is attended, and so on. Because the previous estimate summarizes the impact of the previous cues, and the effect of a new cue is the same regardless of the *cue values* observed for the previously attended cues, the adjustment naturally represents the *independent effect of each cue*. Configural (or multiplicative) effects of different cue values are therefore not naturally represented.[4]

This, of course, is not to deny that people can multiply digits (already most third-graders do), only to hypothesize that this capacity draws not on spontaneous and intuitive cue integration, but on elaboration in working memory of declarative facts from long-term memory. With intuitive controlled judgment, the judge has declarative knowledge of the cues and their relations to the criterion, but not of the linear additive integration rule that "emerges" from the sequential process. In terms of dual systems theories (Darlow & Sloman, 2010), this process occupies a middle ground in that the process is controlled, the cues are attended, and easily verbalized (as in analytic thought), but the integration rule is implicit and not easily verbalized (as with intuition). Because this is the process we turn to when the answer cannot be retrieved or deduced from analytic principles, it has aptly been referred to as "... *a cognitive activity of last resort*" (p. 139, Adelman, Stewart, & Hammond, 1975).

Consider such an updating that involves estimating a probability $p(C)$, based on sequential consideration of two probabilities, $p(A)$ and $p(B)$. If the normative integration rule is additive; $p(C) = p(A) + p(B)$ (e.g., if C is the disjunction of the exclusive events A and B), sequential adjustment can implement the normative rule. For example, adding .1 after observing $p(B) = .1$ produces the correct result regardless of whether the previously attended $p(A)$ is .5 or is .9. In the case of the multiplicative rule in Eq. (2), $p(C) = p(A) \cdot p(B)$, the *same* adjustment in response to $p(B)$, independently of $p(A)$, cannot implement the rule. If $p(A)$ is .9., the adjustment implied by $p(B)$ is to subtract .81 to yield .09; if $p(A)$ is .5, the adjustment is to subtract .45 to yield .05. Thus, sequential adjustment of the independent effect of each cue is ill suited to capture many of the rules of probability theory (Juslin et al., 2009).

[4] Note that independence in this sense will then hold regardless of whether the sequential adjustment process implements a simple summation model (i.e., adding up the inputs) or a weighted average model, as long as the effect of a specific cue at a certain time, in the form of an adjustment from a previous estimate, is the same regardless of the values of the cues processed previously in this process of sequential adjustment.

In belief revision tasks, where the belief is repeatedly updated in the face of new evidence (rather than just once as in a base-rate problem), it has long been known that rather than relying on Bayesian integration people successively average the "old" and "new" data (Hogarth & Einhorn, 1992; Lopes, 1985, 1987; McKenzie, 1994; Shanteau, 1970, 1972, 1975). In regard to the diagnosis problem in the Introduction, the participants may first consider the hit-rate, producing an initial estimate of .96. Thereafter, they consider the base-rate (.02), adjusting the initial estimate in the direction implied by the base-rate (here downwards), the size of the adjustment depending on the perceived importance of the base-rate. Finally, this second estimate is adjusted in the face of the false-alarm rate (.08), as a function of its perceived importance. This will not implement Bayes' theorem, but some "normative insight" can nonetheless be indicated by appreciating that base-rate and hit-rate co-vary positively, but false-alarm rate negatively, with the posterior probability.

On this view, the judgment is not based on a well-defined, generally applicable "heuristic", but on the flexible linear weighting of different information, depending on the judge's causal models, the contextual information, or on the previous feedback (Ajzen, 1977; Birnbaum & Mellers, 1983; Fischhoff, Slovic, & Lichtenstein, 1979; Gigerenzer, Hell, & Blank, 1988; Tversky & Kahneman, 1982). If mental number crunching of Bayes' theorem in its probability version (Eq. (1)) is beyond the ability of most people (Gigerenzer & Hoffrage, 1995), it follows that instruction and training cannot install normative integration in this task but—at best—optimize the weights used in an intuitive additive approximation.

Applied to Linda, people may be able to assess the probability that Linda is feminist (high) or a bank teller (low). Because they know of no feminist bank tellers, they infer the conjunctive probability by combining the known probabilities. To the extent that they do not know, or fail to retrieve, the multiplicative rule from probability theory, they are likely to fall back on intuitive linear additive integration. Because a weighted average falls between the two probabilities, the result is a conjunction error. Representativeness *may* be involved in assessment of the constituent probabilities, but it's not the *cause* of the fallacy; the cause is the use of weighting and adding rather than multiplicative probability integration. In other words, even if the constituent probabilities are estimated in other ways, such as relative frequencies, linear additive integration will produce a conjunction fallacy.

In Juslin et al. (2009) we provide an extensive review of the literature, which supports the claim that both base-rate neglect and the conjunction fallacy may be caused not primarily by use of the representativeness heuristic per se, but by linear additive integration of information. Nilsson et al. (2009) moreover report a series of new experiments that test this explanation in regard to the conjunction fallacy. The experiments reported below complement these previous studies by testing the hypothesis that these cognitive biases should diminish if the problems are presented in a format that allows addition rather than multiplication.

3.3. Exemplar memory judgment

A third possibility is that the judgment is based on memory of previous concrete judgment cases (exemplars; Medin & Schaffer, 1978; Nosofsky & Johansen, 2000). For example, if a participant has previously encountered a similar medical diagnosis task (e.g., low base-rate, relatively high hit-rate, relatively low false alarm rate), the correct answer to the previous problem is a basis for guessing the answer to the new problem (i.e., "Presumably a surprisingly low probability also in this problem!"). Exemplar models predict that performance should be better for repeated old problems seen in training than for new problems, and performance should be good within the training range (interpolation), but poor for extrapolation outside of the training range (see DeLosh, Busemeyer, and McDaniel (1997) and Juslin, Olsson, and Olsson (2003) for discussions). We return to exemplar memory in the context of Experiment 2 that involves feedback training with base-rate problems.

3.4. Overview of the experiments

In the following, Experiments 1 and 2 address base-rate neglect with the medial diagnosis task, demonstrating that people are inclined to use linear additive integration both before and after instruction (Experiment 1) or extensive feedback training (Experiment 2), but that base-rate neglect is diminished and easily rectified by an additive format in log odds. In Experiment 3, we extend the conclusions from the first experiments by showing that also the conjunction fallacy can be nominally eliminated by an additive logarithm format.

4. Experiment 1: Nominal elimination of base-rate neglect

From the framework in Fig. 1 it follows that if people cannot retrieve and use analytical principles or answers to similar concrete problems, an important constraint on their ability to reason with probability is the spontaneous inclination for linear additive integration with intuitive judgment. Is there any way in which we can test this conjecture?

One of the basic laws of logarithms implies that the product,

$$A = B \cdot C, \tag{4}$$

implies the following additive relationship in logarithms,

$$\log(A) = \log(B) + \log(C). \tag{5}$$

Eqs. (4) and (5) essentially state the same fact, but express it in two different metrics, one that requires multiplication and one that requires addition of the constituents. One metric with the virtue of transforming Bayes' theorem into additive form is the *logs odds ratio format*:

$$\log\left(\frac{p(D|P)}{p(\bar{D}|P)}\right) = \log\left(\frac{p(D)}{p(\bar{D})}\right) + \log\left(\frac{p(P|D)}{p(P|\bar{D})}\right). \tag{6}$$

Positive log odds ratios favor the hypothesis, zero corresponds to ambiguity and negative log odds ratios speak against the hypothesis. Log odds run on the interval $[-\infty, \infty]$, where two examples of natural anchors back to the more familiar odds ratio format is log odds ratio 1, implying that the hypothesis (D) is ten times as likely as its negation, and the log odds ratio -1, implying that the negation ($\bar{D}$) is ten times as likely as the hypothesis. The log prior odds in the above medical diagnosis task is -1.69, the log likelihood ratio is 1.08; thus the posterior odds ratio is $-.61$, speaking distinctly against that the patient has the disease.

A first aim of Experiment 1 was to test the hypothesis that performance in base-rate tasks is improved by logarithm formats that require addition rather than multiplication. If this hypothesis is correct, then more base-rate neglect should be observed when problems use a metric that requires multiplication (probability in Eq. (1) or odds in Eq. (2)) than when the problems use a metric that requires linear additive integration (log odds in Eq. (6)). The purpose here is *not* to advocate this log measure of uncertainty as a practical method to "debias" human judgments (although we return to this possibility in Section 7). Rather, it serves as a vehicle for testing the hypothesis that base-rate neglect derives at least in part from an inability to integrate the information in the normative and multiplicative manner.[5] A secondary aim was to validate the assumption that people are spontaneously inclined to rely on linear additive integration in the probability version of the base-rate problem.

In Experiment 1, participants assessed the posterior probability for the same 18 problems in each of three formats, the *probability format* (Eq. (1)), the *odds format* (Eq. (2)), and the *log odds format* (Eq. (6)). The probability format requires integration of three probabilities (base-rate, hit-rate, and false-alarm rate), while the other two formats only require integration of information about the prior odds and the likelihood ratio. The odds and the log odds format are thus equally complex in this respect and only differ in the integration rule. Half of the participants only received a *Metric instruction*, which introduced the metric (probability, odds, and log odds) without information about Bayes' theorem. This instruction explained the interval on which the metric runs and provided a few anchor points for its interpretation (e.g., the sentences that follow Eq. (6)). The other half received a *Computational instruction*, which in addition, also explicitly presented and explained Bayes' theorem in its relevant form (i.e., Eqs. (1), (2), (6), depending on the format), along with a concrete computational example.

Prediction 1 was that performance with log odds should be better than with probability and odds already without computational instructions, because the normative integration with log odds is additive, consistently with intuitive integration. The most likely naïve default might, however, be a weighted mean rather than a sum, as suggested by numerous studies that address other judgment contents (e.g., Anderson, 1996; Hogarth & Einhorn, 1992; Juslin et al., 2008; Lopes, 1985, 1987; Roussel et al., 2002; Shanteau, 1970, 1972, 1975), and there is thus little reason to expect *perfect* performance with metric instructions and log odds.

Prediction 2 was that people should find it easier to improve performance with computational instructions and log odds format than with the other two formats, because with the log odds format the normative response is invited both by the intuitive and the analytic route in Fig. 1. With probability format the evidence suggests that people have difficulty with mentally performing the computations required (Gigerenzer & Hoffrage, 1995). If this derives from cognitive constraints (Juslin et al., 2008), it is not rectified by instructions.

Prediction 3 was that with probability format judgments should be better fitted by a linear additive model than by a multiplicative model (with Bayeś theorem as a special case). Most participants are unlikely to spontaneously retrieve knowledge of Bayes' theorem (which likely is unknown to most of the participating undergraduates), and even if they do, they will have extreme difficulty with mentally performing these computations (Gigerenzer & Hoffrage, 1995). With metric instruction we expect the participants to rely on some simple non-integrative strategy, like reporting the hit-rate (the "Fisherian algorithm", Gigerenzer and Hoffrage), or some rudimentary, idiosyncratic linear additive integration, reflecting, at least partial, understanding that also the base-rate and (or) the false alarm rate are relevant.

Prediction 4 is that also with computational instruction, and thus explicit knowledge about Bayes' theorem, the participants should be unable to implement Bayes' theorem. People should find it easy to improve their judgments by adapting their linear weighting to approximate Bayes' theorem, but they should be reluctant or unable to shift to multiplication. The instruction should accordingly not result in the best fit for a multiplicative model, but in the best fit of a linear additive model with weights converging on those that allow approximation of Bayes' theorem. Assuming that multiplication of two digits according to a rule is within the working memory constraints of most participants, Fig. 1 is more ambiguous with regard to the cognitive processes with the odds format. With the metric instructions it seems reasonable to expect that the participants should fall back on the default of intuitive linear integration. With computational instruction, however, they could either continue with this intuitive integration or use the analytic process implied by the instruction (e.g., use the analytic rule that "normative integration is to multiply prior and likelihood odds", and if these digits are 3 and 7, retrieve that "$3 \times 7 = 21$"). In contrast to the probability format, analytic number crunching seems to be well in reach with the odds format.

[5] As further discussed in Section 7, we refer to the decrease or elimination of the observed biases reported in this article as "nominal", so signify that they need not imply a deep conceptual understanding of the probability metrics or an ability to generatively generalize this understanding to novel tasks. These effects are merely the predictions if people spontaneously and naively try to integrate information linearly and additively.

4.1. Methods

4.1.1. Participants

Thirty undergraduate students participated (18 female and 12 male; average age = 25). As compensation, participants received either course credits or a movie ticket.

4.1.2. Apparatus and materials

Stimuli and instructions were included in booklets.

The base-rate problems concerned unspecified diseases (one for each problem) and included information concerning the *base-rate* of the disease in the Swedish population, the *hit rate* of the test, and the *false-alarm rate* of the test. The base-rate was randomly sampled from a uniform distribution between 0 and .5 (sample mean .236 across the 18 problems), the hit-rate was randomly sampled from a uniform distribution between .5 and 1 (sample mean .742), and the false-alarm rate was randomly sampled from a uniform distribution between 0 and .5 (sample mean .240). That is, the problems were constrained to involve relatively rare diseases, relatively high hit-rates, and relatively low false-alarm rates, as in a diagnostic test.

There were three versions of each problem, differing only in the way the base-rate, the hit rate, and the false alarm rate were described. The probability (percentage) version was:

> The probability that a person randomly sampled from the population of all Swedes has the disease is 1%. The probability of a positive test result if you have the disease is 90%. The probability of a positive test result if you do not have the disease is 30%.
> What is the probability that a randomly sampled person from the population that receives a positive test result has the disease? _____%

The odds ratio version presented a prior odds ratio and a likelihood ratio:

> The prior odds ratio that a person randomly sampled from the population of all Swedes has the disease is 1/100. The likelihood odds ratio of obtaining a positive result if you have the disease, relative to if you do not have the disease is 3/1.
> What is the odds ratio that a randomly sampled person from the population that receives a positive test result has the disease? _____

The log odds ratio presented a log prior odds ratio and a log likelihood ratio:

> The log prior odds ratio that a person randomly sampled from the population of all Swedes has the disease is −2. The log likelihood odds ratio of obtaining a positive result if you have the disease, relative to if you do not have the disease is .48.
> What is the log odds ratio that a randomly sampled person from the population that receives a positive test result has the disease? _____

Formally, there accordingly were 18 base-rate problems, each appearing once as a probability version, once as an odds version, and once as a log odds version. We assumed that independently of version, problems that included numbers with few decimals and/or even numbers such as .50 and 1.50 rather than .53 and 1.48 would be easier to solve. To control for this aspect, six problems including "simple" numbers were created for each format (probability, odds, and log odds). Hence, among the 18 problems there were six that produced probability versions with "simple" numbers, six that produced odds versions with "simple" numbers, and six that produced log odds versions with "simple" numbers.

In regard to each of the three formats (probability, odds, log odds), the participants either received a metric or a computational instruction. Each of the three *metric instructions* briefly explained one of the metrics by describing the formal relationship between the metric and probability (for odds ratios and log odds ratios) and by translating three salient reference points back to the more familiar probability format and into their approximate verbal meaning (i.e., "very unlikely", "even chance", and "very likely"). In addition, the direction of the metric was explained. The instruction with the probability format thus (and probably redundantly) explained the meaning of "1%", "50%", and "99%" probability and that, the more the probability exceeded 50%, the higher the probability of the hypothesis relative to its negation and the more the probability fell below 50%, the higher the probability of the negation relative to the hypothesis. The metric instruction for odds ratios explained the meaning of odds ratios "1/100", "1/1", and "100/1", by translating them into their approximate probabilities. It was stated that the more the odds ratio exceeded 1/1, the higher the probability of the hypothesis relative to its negation and the more the odds ratio fell below 1/1, the higher the probability of the negation relative to the hypothesis.

The metric instruction for log odds ratios explained the meaning of log odds ratios "−2", "0", and "2", by translating them into their corresponding odds ratios and approximate probabilities. It was stated that the more the log odds ratio exceeded 0, the higher the probability of the hypothesis relative to its negation and the more the log odds ratio fell below 0, the higher the probability of the negation relative to the hypothesis.

In each case, it was emphasized that the reference points mentioned were just intended as examples and that the assessed value could take any value admissible with the metric in question (i.e., identified as $[0, 1]$, $[0, \infty]$, and $[-\infty, \infty]$, depending on the condition). To ensure that the participants had some grasp of the metric in question, the three instructions also included five simple multiple choice questions that required the participants to translate the metric in question back to the more familiar probability format. For example, the odds ratio metric instructions included a question on whether an odds ratio of 1 indicated that the likelihood of disease given a positive test result was high, low, or intermediate. If the five questions were not correctly answered, the participants were asked to read the instruction once

more. Virtually all participants answered these simple questions correctly the first time (i.e., "simple" in the sense that the answers were explicitly stated on the previous page of the instruction, which they were reading), and all got it right the second time.

The three *computational instructions* included the metric instruction plus an in depth instruction on how to use Bayes' theorem to solve the probability versions (the probability computational instruction), the odds ratio versions (the odds ratio computational instruction), or the log odds ratio versions (the log odds ratio computational instruction). The instructions explained the meaning of the components of the problem (e.g., the base-rate, the hit rate, & the false alarm rate in the probability versions) and provided a numerical example of how they are combined in Bayes' theorem (as modified to suit the metric in question). In all conditions, the instructions remained available to the participants throughout the condition in question. The participants were not allowed to make computations with paper and pencil.

There were two types of booklets. The metric booklet included 54 problems, 18 probability, 18 odds ratio, and 18 log odds ratio problems. Each set of 18 problems was preceded by the relevant metric instructions. The computational booklet included the same 54 problems, but each set of 18 problems was preceded by a relevant computational instruction.

4.1.3. Design and procedure

The experiment was a 2∗3 mixed design with instructions (metric vs. computational; between subjects) and version (probability, odds ratio, and log odds ratio; within subject) as independent variables. All participants were tested individually. At arrival, participants were handed a booklet. Half of the participants were handed the metric booklet and the other half were handed the computational booklet. All participants were asked to solve all 54 problems. One third of the participants were asked to start with the probability versions and end with the log odds ratio versions, one third to start with the odds ratio versions and end with the probability versions, and one third to start with the log odds ratio versions and end with the odds ratio versions. The experiment lasted 20–40 min.

4.2. Results

We addressed the data by two kinds of analysis. First, we analyzed performance with Mean Absolute Error (MAE) from the "correct" posterior probability and report the observed bias relative to Bayes' theorem. We analyze performance measures collapsed across both of the instruction conditions, then separately for the Metric and the Computational instructions. Before analysis, for the sake of comparability the data in the odds and log odds conditions were transformed into probability. Second, because a key-hypothesis is that in the original (probability) base-rate tasks the participants spontaneously rely on linear additive integration, we compare a multiplicative and a linear additive model with respect to these data.

4.3. Performance: effects of format and instruction

Collapsed across both of the instruction conditions, MAE was higher (poorer) with the probability version than with the other two versions (both $p < .01$ by Wilcoxon test). The odds version produced significantly higher (poorer) MAE than the log odds version (Wilcoxon; $T = 32$, $Z = 4.13$, $p < .001$). As illustrated in Fig. 2, for all versions MAE decreased with computational instruction (Mann–Whitney; probability: $U = 26$, $Z = 3.57$, $p < .001$; odds: $U = 27$, $Z = 3.53$, $p < .001$; log odds: $U = 13$, $Z = 4.11$, $p < .001$). As predicted, collapsed across instruction conditions, performance was best with logs odds and with all three assessment formats the performance improved significantly with computational instructions.

4.4. Performance: metric instructions

MAE in the metric instruction conditions are presented in Fig. 2A, also showing the lowest (best) MAE that is attainable with optimal linear weighting of the statistics stated in the problem. This best linear MAE was based on regression models with unbounded linear coefficients and no intercept based on the 18 problems, where the digits stated in the problem (in a given format) are independent variables and the correct response (in the same format) as the dependent variable. With metric instructions the participants were briefly introduced to the metric involved (probability, odds ratio, and log odds ratio), but they received no information on how the components should normatively be integrated into a posterior probability.

The probability versions produced significantly higher (poorer) MAE than the odds versions (Wilcoxon; $T = 25$, $Z = 1.99$, $p = .034$) and the log odds versions (Wilcoxon; $T = 6$, $Z = 3.07$, $p = .002$). The odds format produced significantly higher (poorer) MAE than the log odds format (Wilcoxon; $T = 20$, $Z = 2.27$, $p = .023$). Performance with probability and log odds is distinctly poorer than the best performance allowed by linear additive integration (the dotted line in Fig. 2A). Performance with odds is slightly better than expected from optimal linear additive integration, weakly suggestive of multiplicative integration.

Across the 18 problems, the median hit-rate was .75 and the median posterior probability was .40. If the participants ignore the base-rates and only use the hit-rates the median assessed posterior probability should be close to .75, while if they implemented the Bayesian solution their median assessed posterior probability should be close to .40. The median of the median individual judgment across the 18 problems are presented in Fig. 2B. In Fig. 2B, we see that the "base-rate neglect" with probability format (a median judgment of .70) has almost disappeared with the log odds format (.47).[6] The percentage of problems

[6] As expected, the deviations from the normative response observed with metric instruction and log odds format were suggestive of most participants initially addressing the task with something like a weighted mean rather than a sum, an error that was rectified when the participants were given computational instructions.

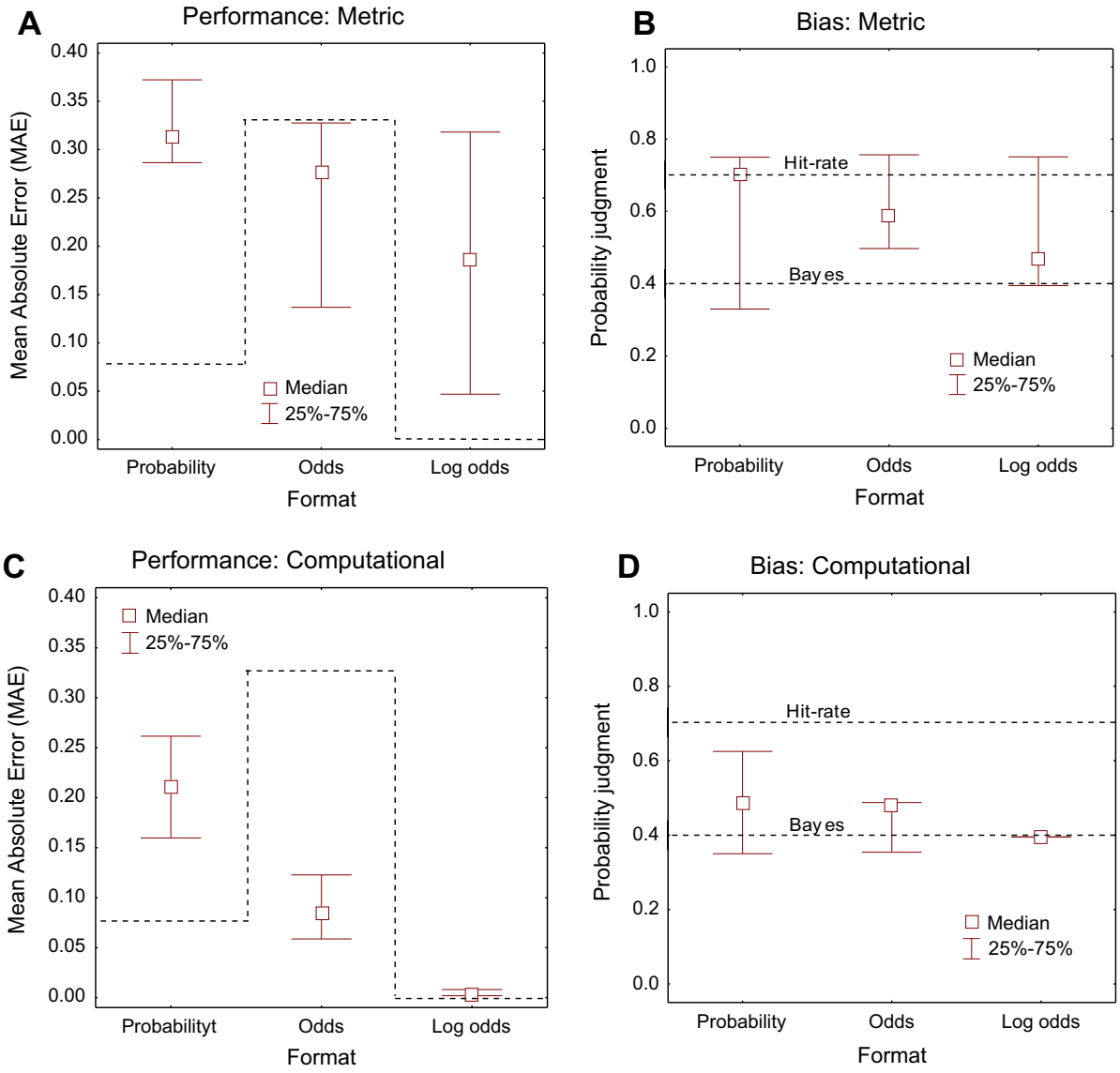

Fig. 2. Experiment 1: Panel A: The median Mean Absolute Error (MAE) between judgment and posterior probability with interquartile index in the Metric instruction condition; Panel B: The median judgment with interquartile index in the Metric instruction condition; Panel C: The median MAE between judgment and posterior probability with interquartile index in the computational instruction condition; Panel D: The median judgment with interquartile index in the Computational instruction condition. In Panels A and C the dotted line represents the lowest MAE achievable by linear additive integration of the figures stated in the problems (see the main text for further explanation). In Panels B and D "Hit-rate" refers to the median predicted probability judgment if the participants respond only to the hit-rate, and "Bayes" to the probability judgment predicted if the participants use Bayes' theorem.

where the responses were exactly correct (rounded to two decimals in probability format) was 1% for the probability, 20% for the odds, and 21% for the log odds format.

4.5. Performance: computational instructions

The MAE in the Computational instruction condition is presented in Fig. 2C. Computational instructions in addition explained how the components should be integrated into a posterior probability according to Bayes' theorem (by multiplication with probabilities and odds, by summation with log odds), and this was concretely illustrated with a computational example. The probability format produced significantly higher (poorer) MAE than the other formats ($p < .01$ by Wilcoxon test). The odds format produced significantly higher (poorer) MAE than the log odds format (Wilcoxon; $T = 15$, $Z = 3.41$, $p = .001$).

As predicted, with a probability format performance was distinctly poorer than the level defined by optimized linear weights (the dotted lines in Fig. 2C). With the odds format, the MAE falls in between the MAE predicted by optimal linear weighting (dotted line) and the MAE implied by normative integration (MAE = 0), suggesting partial, but only partial, ability to implement the multiplicative rule given in the Computational instruction. With the log odds format performance is virtually perfect. Fig. 2D illustrates that there is still some "base-rate neglect" with the other formats, but it has been eliminated with logs odds (median .40). The responses were exactly correct (rounded to two decimals in probability format) for 3% of the probability

versions, 53% of the odds versions, and 85% for the log odds versions.

4.6. Model fit for the probability format

Two regression models were fitted to the judgments $\hat{p}_{po}$ of posterior probabilities in the condition with probability format. The first was a weighted multiplicative model,

$$\log\left(\frac{\hat{p}_{\mathrm{po}}}{1-\hat{p}_{\mathrm{po}}}\right) = a + b_{\mathrm{PR}} \cdot \log\left(\frac{p_{\mathrm{pr}}}{1-p_{\mathrm{pr}}}\right) + b_{\mathrm{LR}} \cdot \log\left(\frac{p_{\mathrm{hr}}}{p_{\mathrm{fa}}}\right) + \varepsilon, \quad (7)$$

where a is an intercept, b_{PR} is the weight assigned to the prior odds ratio and b_{LR} is the weight attached to the likelihood ratio. The special case of this multiplicative model with parameters $a = 0$, $b_{\mathrm{PR}} = 1$, and $b_{\mathrm{LR}} = 1$ is Bayes' theorem in log odds ratio form. If participants are able to perform multiplicative integration and approximate Bayes' theorem, they should be well described by Eq. (7). Note, however, that the free parameters allow the model also to capture various aberrations and idiosyncrasies in a process of multiplicative integration.

The second model is the linear additive model based directly on the three stated probabilities, but with fitted linear coefficients,

$$\hat{p}_{\mathrm{po}} = b_{\mathrm{pr}} \cdot p_{\mathrm{pr}} + b_{\mathrm{hr}} \cdot p_{\mathrm{hr}} + b_{\mathrm{fa}} \cdot p_{\mathrm{fa}} + \varepsilon, \quad (8)$$

where b_{pr} is the weight assigned to the stated base-rate p_{pr}, b_{hr} is the weight assigned to the stated hit-rate p_{hr}, and b_{fa} is the weight assigned to the stated false-alarm rate p_{fa}. In order to limit the number of free parameters and to make the two models as comparable in flexibility as possible, in Eq. (8) we refrained from introducing an intercept.[7] Eq. (7) and (8) thus define two regression models, each with three free parameters fitted to the judgment data.

To evaluate the fit of the models and to investigate average linear parameters both models were fitted separately for each individual participant. The linear additive model provides better fit[8] than the multiplicative model across both instruction conditions (median R for the linear additive model = .97; median R for the multiplicative model = .75: Wilcoxon; $T = 0$, $Z = 4.78$, $p < .001$), as well as separately in the Metric condition (median R for the linear additive model = .99; median R for the multiplicative model = .63: Wilcoxon; $T = 0$, $Z = 3.41$, $p < .001$) and the Computational condition (median *R for the linear additive model = .95; median R for the multiplicative model = .77: Wilcoxon; T = 23, Z = 3.41, p = <.001).*[9]

The median number of significant predictors was 2, suggesting that participants typically integrated several cues. There were however large individual differences and for 40% of the participants only one predictor was statistically significant beyond .05 (although the interpretation of non-significance is complicated by the limited power in the analysis of the individual participants). Among participants with only one significant predictor, there were four participants (13% of all participants) that only used the hit-rate (i.e., a "Fisherian algorithm", Gigerenzer & Hoffrage, 1995). In the Metric condition, on average 35% of the judgments were numerically identical either to the stated hit-rate, base-rate, or false-alarm rate, thus suggesting a non-integrating strategy merely reporting one of the digits. In the Computational condition this percentage decreased to 6% (Mann–Whitney: $U = 54$, $Z = 2.41$, $p = 016$).

As shown in Fig. 2A and C, performance was better in the computational probability condition than in the metric probability condition. Fig. 3, which presents the average parameters of the linear additive model fitted to individual data (95% confidence intervals), illustrates how this improvement was obtained. In the Metric condition, without instruction about Bayes' theorem, on average the judgment was close to a weighted mean of the base-rate and the hit-rate, with the hit-rate receiving most weight. However, on average they do not assign a non-zero weight to the false-alarm rate. As we have seen, the computational instruction informing the participants about Bayes' theorem does not seem to instigate a shift to the appropriate multiplicative model. Instead, they adopt linear parameters that allow the linear additive model to better approximate Bayes' theorem, which includes adopting a negative weight for false-alarm rate. In sum: the linear additive model provided better fit than the generalized multiplicative model with both instructions and the improved

[7] Although the constraint not to include an intercept was primarily motivated by the ambition to keep the number of free parameters equal we note that it is consistent with the observation that people often spontaneously use a weighted average (Anderson, 1996), which has been observed also in probability tasks (Nilsson et al., 2009). The fit of the linear additive model can of course only be further improved by adding an intercept, and in that sense the decision is conservative with regard to the observed superiority of the linear additive model.

[8] Because the regression models are applied to prediction of very different metrics, log odds ratios in the multiplicative model and probabilities in the linear additive models, evaluating the fit in terms of prediction error is difficult. We therefore relied on the multiple correlation R to evaluate the fit of the models.

[9] When the regression models are fitted to individual participants, both models have three free parameters that are fitted to 18 data points, which actualizes the potential problem of statistical overfit. In order to validate the conclusions from the regression modeling of individual participants we therefore also analyzed the predictions by two corresponding a priori (parameter-free) models. We computed the correlation between the participant's judgments and Bayes' theorem (Eq. (1)) and the corresponding correlation with the best linear additive approximation to Bayes' theorem. The best linear approximation was estimated for the case where the base-rate is uniformly distributed between 0 and .5, where the hit-rate is uniformly distributed between .5 and 1, and the false-alarm rate is uniformly distributed between 0 and .5, as relevant to our medical diagnosis tasks. The best approximate linear model is, $\hat{p}_{\mathrm{po}} = .251 + 1.303 \cdot p_{\mathrm{pr}} + .252 \cdot p_{\mathrm{hr}} - 1.406 \cdot p_{\mathrm{fa}}$, where p_{pr}, p_{hr}, and p_{fa} are the base-rates, the hit-rates, and the false-alarm rates, respectively, stated in the problems. In the Metric instruction condition the mean correlation with Bayes' theorem was .375 and the mean correlation with the linear additive approximation was .477 (Wilcoxon; $T = 0$, $Z = 3.41$, $p < .001$). In the Computational instruction condition the mean correlation with Bayes' theorem was .656 and the mean correlation with the linear additive approximation was .694 (Wilcoxon; $T = 21$, $Z = 2.22$, $p = .027$). Although the differences were small – not surprising given that the linear model is an approximation to Bayes' theorem ($r = .95$) – the differences consistently favored the linear additive model over Bayes' theorem. Note, however, that the best linear approximation is not a good description of the participants in the Metric condition, which in contrast to the best linear additive approximation give no significant negative weight to false alarms (see Fig. 3).

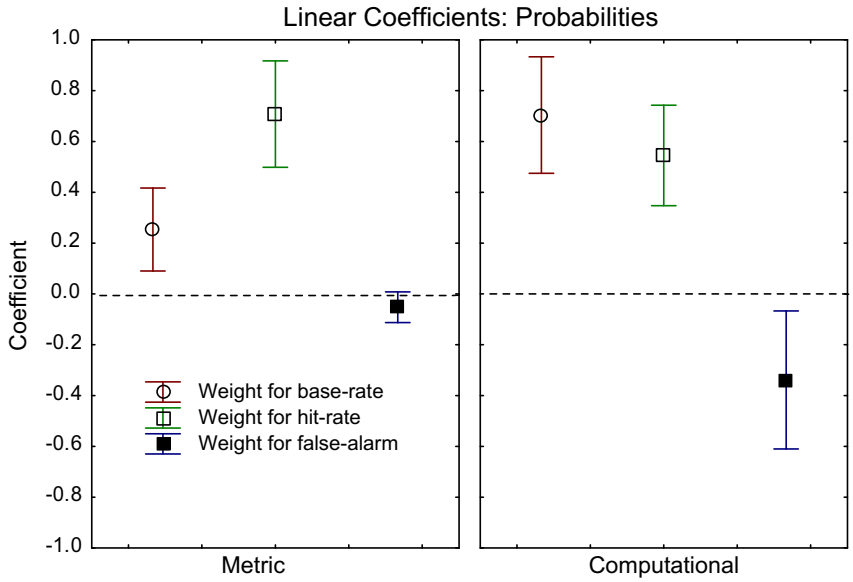

Fig. 3. Mean linear coefficients with 95%. Confidence intervals ($N = 15$) for the linear additive models in the conditions with a probability assessment format.

performance with computational instructions is explained, not by a shift to Bayes' theorem, but by adapting the linear additive model to make it a better approximation of Bayes' theorem.

4.7. Discussion

As predicted, the use of a log odds format immediately improved performance (Prediction 1). Already without computational instructions, performance with the log odds format was significantly better than performance with the other two formats. That the log odds format yielded significantly better performance than the odds format, where both formats only require integration of two numbers, suggests that the integration per se plays a key role for this difference. Moreover, the better performance with odds than probabilities confirms that the analytic number crunching of two digits with odds is within the scope of working memory limitations, in contrast to the multiplication with a probability format. Only with odds, performance exceeded what is allowed by linear additive integration.

Only the log odds format allowed the participants to "perfect" their performance (Prediction 2). When the problems were framed in way that makes Bayes' theorem additive (log odds ratios), just a few minutes of tutoring in the Computational instruction condition was sufficient to nominally eliminate the base-rate neglect in all of the participants. This was in stark contrast to how the computational instructions affected performance with the other two formats. Even after computational instruction, performance with the probability format was no better than it was already with metric instruction on the log odds ratio format.

The participants with probability assessment format were much better fitted by a linear additive model than by a multiplicative model, with Bayes' theorem as a special case (Prediction 3). Most of the participants integrated several cues, typically the hit-rate and the base-rate in the Metric condition, while in addition including the false-alarm rate in the Computational condition. A minority in the Metric condition, however, relied on a non-integrative strategy, responding with the hit-rate ("Fisherian algorithm", Gigerenzer & Hoffrage, 1995).

The computational instructions improved accuracy also with the probability format. However, the instructions did so by helping the participants to adjust their weights rather than by enabling them to use Bayeś theorem (Prediction 4). As predicted, a brief instruction was sufficient for the participants to flexibly shift their weights to better approximate the output of Bayes' theorem. Consistently with the original notion of a base-rate neglect bias, more weight was spontaneously assigned to the hit-rate than to the base-rate. Somewhat surprisingly, the only information that was truly ignored by the participants was the false-alarm rate. Relying on a "Fisherian algorithm" (Gigerenzer & Hoffrage, 1995), committing what has been called "the inverse fallacy" (Villejoubert & Mandel, 2002), implies ignoring both base-rate and false-alarm rate. That people respond both to the base-rate and the hit-rate, but not to the false alarm rate is not captured by these notions. Interestingly, it is consistent with a literature suggesting that people have limited ability to consider the implications of the alternative hypotheses in hypothesis testing, so called "pseudo-diagnosticity" (Ofir, 1988).

5. Experiment 2: Learning Bayes' theorem from feedback

The results from Experiment 1 suggest that people have difficulties in adopting a multiplicative strategy for integrating information even after explicit instructions on how to do so. The aim of Experiment 2 was to investigate if extensive outcome feedback is sufficient to make people

adopt multiplicative information integration in a base-rate task.

As in Experiment 1, participants judged the posterior probability that a person has a disease from a base rate, a hit rate and a false alarm rate (the probability condition), from a prior odds and a likelihood ratio (the odds condition), or from a log prior odds and a log odds ratio (the log odds condition). In Experiment 2 people train with outcome feedback for similar diagnosis problems. Therefore, the third route in Fig. 1, exemplar memory (Medin & Schaffer, 1978; Nosofsky & Johansen, 2000), also becomes a plausible way to make the judgments (i.e., by retrieving the posterior probability of a previous problem with similar base-rate, hit-rate, and false alarm rate). To investigate this possibility, at test participants were required to make judgments of posterior probabilities outside of the training range (i.e., to extrapolate). Exemplar models predict that, because the judgments are a weighted average of the posterior probabilities observed in training, performance should be better for repeated old problems seen in training than for new problems, and performance should be good within the training range (interpolation), but poor for extrapolation outside of the training range (see Delosh et al., 1997; Juslin et al., 2003 for discussions). However, if they rely on information integration based on some abstract rule, such as Bayes' theorem or linear additive integration, which allows the same performance for old, interpolation, and extrapolation problems, there should be no such difference in the performance. Assuming that participants use linear additive integration, we predicted no such old–new differences.

The predictions were otherwise similar to in Experiment 1. Because people spontaneously integrate information additively they will perform better with the log odds format and will find it much easier to learn to make accurate judgments with this format. The participants in the probability condition should continue to use an additive strategy even after extensive feedback, hence performing poorly, while participants in the log odds condition should rapidly learn to make accurate judgments. Data from the probability condition should moreover be better fitted by a linear additive model than by a multiplicative model. The participants in Experiment 2 are given repeated feedback and should, therefore, be better able to optimize the weights in the linear additive model. As a result, we expected that they should be able to improve their performance to the maximum level allowed by linear additive integration, but not be able to improve their performance beyond this level.

5.1. Methods

5.1.1. Participants

Thirty-six undergraduate students at Uppsala University, 17 male and 19 female (average age 24.8 years), received a movie voucher or course credits for participating.

5.1.2. Procedure and material

Tasks, instructions, and procedures were similar to in Experiment 1, with the following qualifications. The participants were randomized into three conditions; the probability condition, the odds condition and the log odds condition. They were given written instructions corresponding to the metric instructions in Experiment 1.The problems were generated by randomly sampling a base-rate from a uniform distribution between 0 and .5, a hit-rate from a uniform distribution between .5 and 1, and a false alarm-rate from a uniform distribution between 0 and .5, as appropriate for a medical diagnosis task. The problems in the odds and log odds conditions were generated by transformation into the relevant format. The problems were presented one by one on the computer screen. After every problem, the participant was presented with the correct posterior probability expressed in the relevant format. In all conditions, the task was to improve the judgments by help of the feedback. In the training phase, participants received 180 trials, and in a test-phase a further 60 trials without feedback. Throughout the training phase the problems were constrained to only produce posterior probabilities within the interval .05–.95. The test-phase consisted of three sorts of problems; 20 old problems that were repeated from the training phase, 20 new (interpolation) problems within the training range, and 20 new problems outside of the training range .05–.95 (i.e., with lower posterior than .05 or higher than .95).

5.2. Results

5.2.1. Learning performance

As in Experiment 1, all data was transformed to the 0–1 probability scale. Performance during the training phase in terms of Mean Absolute Error (MAE) from the posterior probability is summarized in Fig. 4. Dotted horizontal lines represent the minimum MAE achievable with linear additive integration (i.e., the MAE that would have been reached if optimal weights, given the stimuli material used in Experiment 2, had been used). First, there is learning in all three conditions (i.e., MAE decreases over the training phase). Second, participants in the log odds condition reach more or less perfect performance already in the second block (Fig. 4C). Third, while the MAE in the probability condition quickly stabilizes at the minimum level achievable with linear additive integration (Fig. 4C), the MAE in the odds condition stabilizes in between the minimum level achievable with linear additive integration (the dotted line) and normative integration (MAE = 0). As in Experiment 1, this suggests that participants in the probability condition continue to use intuitive additive integration, but that many of the participants in the odds condition were able to discover the multiplicative rule from the feedback, which is implemented with partial success. A Kruskal–Wallis ANOVA by ranks on the judgments made in the training phase showed a significant effect of condition on MAE ($H = 14.15$, $n = 36$, $p < .001$). The log odds condition has a smaller (better) median MAE ($Md = .023$) than the odds ($Md = .051$) and probability ($Md = .138$) conditions. Testing the pair-wise differences with three Mann–Whitney U-tests revealed all pair-wise differences to be significant (all $p < .05$).

5.2.2. Test performance

A Kruskal-test based on the MAE from the test phase revealed a significant effect ($H = 19.60$, $n = 36$, $p < .001$) with

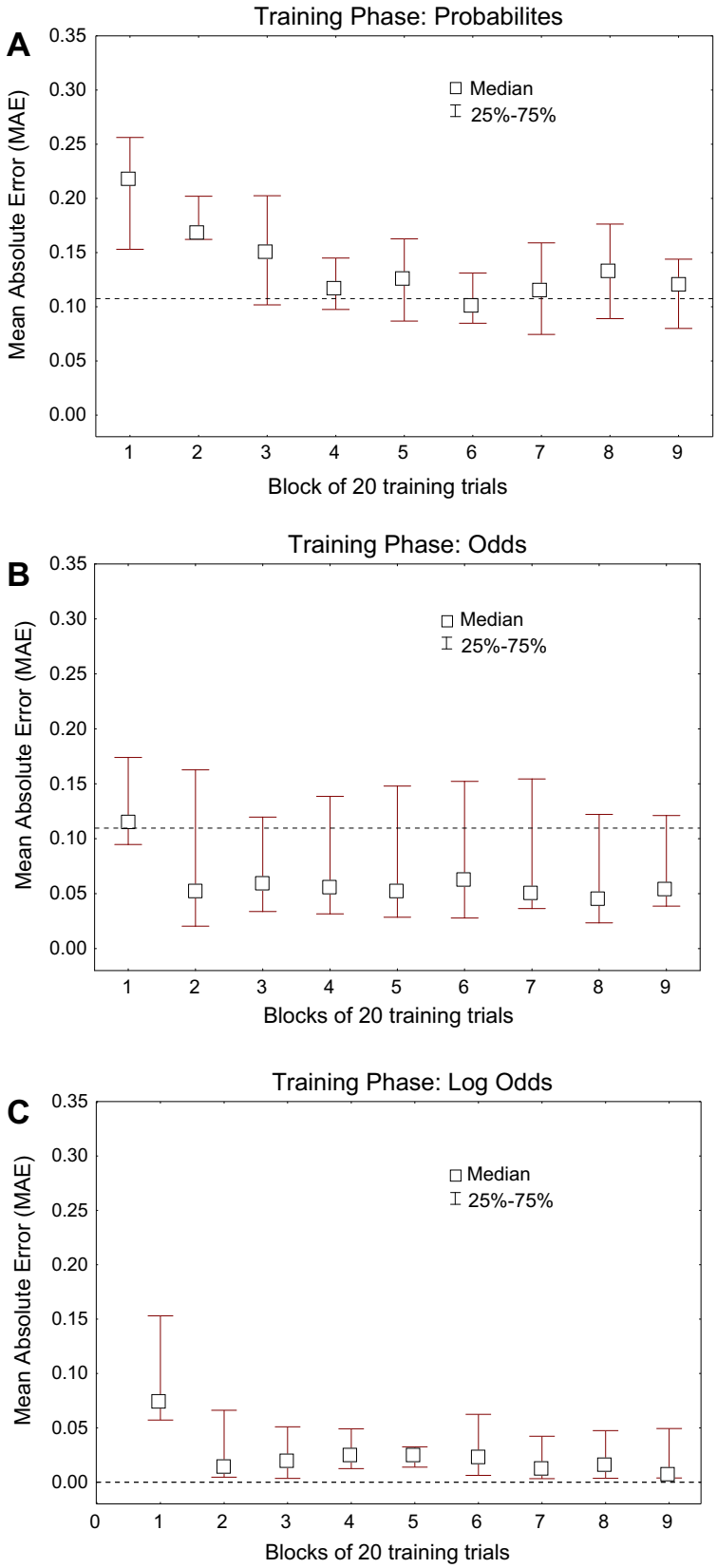

Fig. 4. Performance in the training phase of Experiment 2 in the three conditions expressed as median Mean Absolute Error (MAE) with interquartile index for blocks of 20 trials. Panel A: The probability condition. Panel B: The odds condition. Panel C: The log odds condition. The dotted line in each panel represents the lowest MAE achievable by linear additive integration of the figures stated in the problems (see the main text for further explanation).

the log odds condition having smaller (better) median MAE (*Md* = .012) than the odds (*Md* = .021) and the probability (*Md* = .137) conditions (Fig. 5A). Mann–Whitney *U*-tests revealed that all three pair-wise differences were statistically

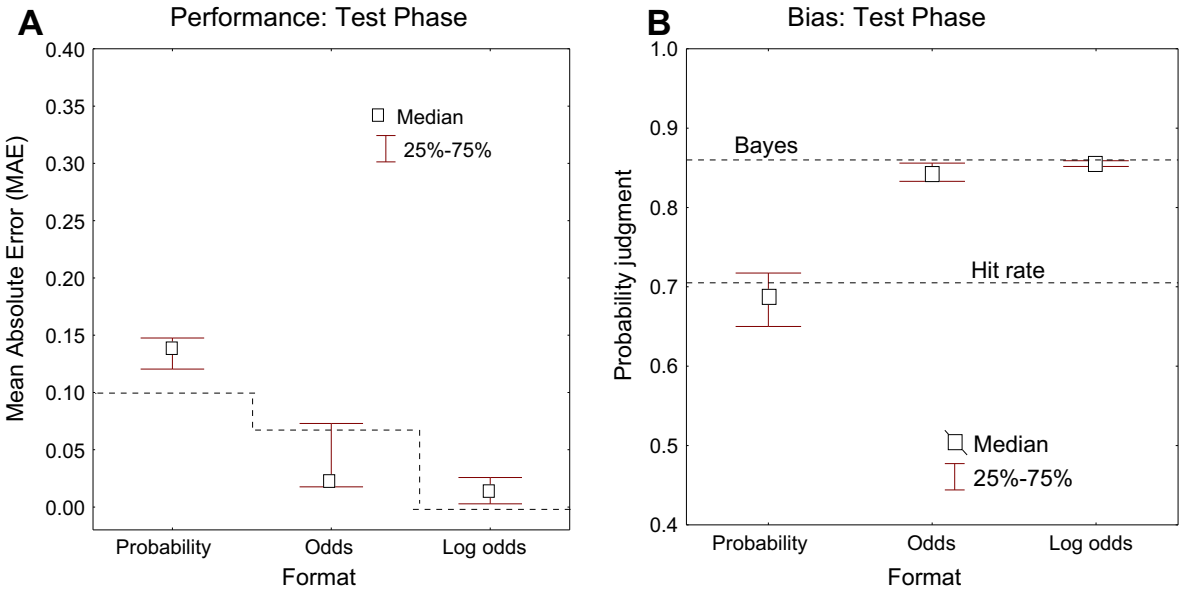

Fig. 5. Performance during the test phase of Experiment 2 in terms of Mean Absolute Error (MAE) and the base rate neglect bias. Panel A: the median MAE with interquartile index in the test phase. Panel B: Median judgment with interquartile index in the test phase. In Panels A the dotted line represents the lowest MAE achievable by linear additive integration of the figures stated in the problems. In Panel B "Hit-rate" refers to the median predicted probability judgment if the participants respond only to the hit-rate, and "Bayes" to the median probability judgment predicted if they use Bayes' theorem.

significant (all $p < .05$). Because in this experiment the base-rate was sampled from the entire interval [0, 1] and extrapolation items required extreme posterior probabilities, the median posterior in the test phase of is actually higher than the hit-rate (see Fig. 5B), in contrast to what has been common in most previous experiments (and in Experiment 1 of this article). A useful property of this data set is that it demonstrates that the improvement with log odds is not merely the result of a general tendency to produce less extreme judgments (as might be argued in Experiment 1), but that it truly tracks the posterior probabilities.

During the test phase the participants made judgments on items that they had experienced during training (old) and items that that they had not experienced during training (new). A Wilcoxon matched pairs test showed no significant difference in the MAE between new and old items ($T = 303$, $Z = .47$, $p = .64$). The test items were either within the range of the items in the training phase (interpolation) or outside of the range of the training items (extrapolation). A Wilcoxon matched pairs test showed no significant difference in the MAE between interpolation and extrapolation items ($T = 292$, $Z = .64$, $p = .52$). As illustrated in Fig. 5B, in the test phase participants in the Log odds condition gave judgments very close to what would be expected from Bayes' theorem, while the judgments in the probability format were closer to the stated hit rate. The percentage of problems where the responses were exactly correct (rounded to two decimals in probability format) was 4% for the probability, 35% for the odds, and 73% for the log odds format.

Model Fit The models described in Eq. (7) and (8) were fitted to the test phase data in the probability condition. When fitted individually for each participant the linear additive model provided better fit than a multiplicative model (median R for the linear additive model = .98, median R for the multiplicative model = .86: Wilcoxon; $T = 0$, $Z = 3.06$, $p = 0.002$). Among the 12 participants in the probability condition 10 (83%) had significant beta-weights for all three components, one (8%) had two significant predictors and 1 (8%) had one significant predictor. On average in training 5.2% of the judgments were numerically identical either to the hit-rate, base-rate, or false-alarm rate, suggesting a non-integrating strategy of reporting one of the stated digits. In the test phase, the percentage was 3.8 and all came from two participants. Most participants apparently integrated the cues into a judgment.

In Experiment 1, the "naïve" participants in the Metric condition that received no instruction on Bayes' theorem typically weighted the base-rate and the hit-rate in their judgments, but ignored the false-alarm rate. To explore these patterns further, the linear additive model (Eq. (8)) was fitted to individual data from the learning phase of Experiment 2. In this analysis, the learning phase was divided into six blocks of 30 trials and the model was fitted separately for each block. The results are shown in Fig. 6. The best fitting coefficients for the linear additive model in the first block replicated the data from the Metric condition in Experiment 1, responding significantly to the base-rate and the hit-rate, but ignoring the false alarm rate. With feedback the participants however adopted the weights to better approximate Bayes' theorem, which lead to the improved performance observed in Fig. 4A.

5.3. Discussion

In Experiment 2, participants in the probability condition performed relatively poorly, even after extensive feedback. Interestingly, their performance did improve as a function of training, but never exceeded the maximum level that

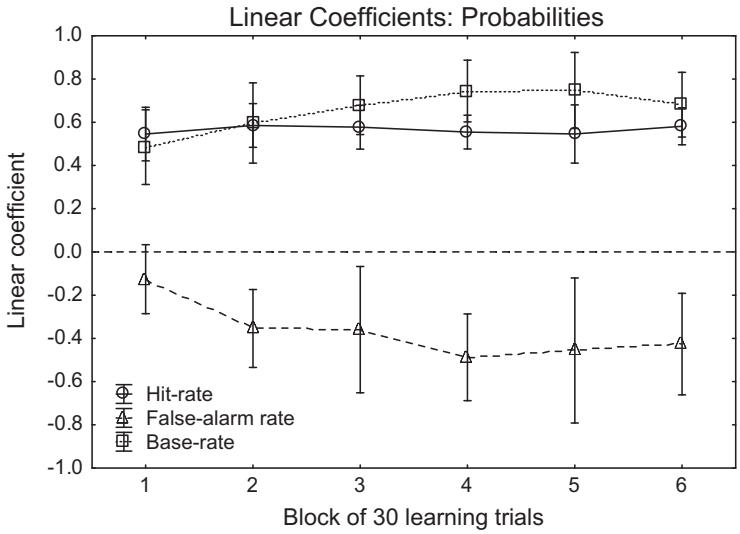

Fig. 6. Mean linear coefficients with 95% confidence intervals ($N = 10$) from the best-fitting linear additive model for base-rate, hit-rate and false alarm-rate in the probability condition during the training phase for the individual participants in Experiment 2.

was obtainable with the linear additive model. Hence it seems that feedback is not sufficient for participants to learn to make judgments that are in accordance with Bayes' theorem when the information is presented in a probability format. However, when information was presented in a format requiring additive integration, that is, the log odds format, feedback rapidly enabled perfect performance.

In addition, the results of Experiment 2 show that participants in the probability condition use a judgment strategy that is best described by a linear additive model and that they do so even after having had extensive feedback on their judgments. Experiment 1 showed that participants had difficulties with using Bayes' theorem even after explicit instructions on how to do so (Computational instruction). The results from Experiment 2 extend this finding to outcome feedback. That is, even with extensive feedback participants have difficulties with integrating probabilities according to Bayes' theorem. There was no evidence for old–new differences that would suggest exemplar memory judgments.

As shown in Fig. 4, the participants in the odds condition performed clearly better than what could be expected from linear additive integration, suggesting that they were able to induce the normative multiplicative rule from the outcome feedback and, and least in part, to successfully implement it. These results essentially replicate those from the Computational condition in Experiment 1. However, as shown by the difference in performance between the odds condition and the log odds condition, the two conditions where the participants were required to integrate two pieces of information, it was substantially easier to detect and implement an additive rule than to detect and implement a multiplicative rule.

6. Experiment 3: Nominal elimination of the conjunction fallacy

Experiments 1 and 2 suggest that participants integrate the components in the probability version of the medical diagnosis task by linear additive integration, and that this is a key constraint on their ability to compute the correct posterior probability. In contrast, a few minutes of tutoring on how to solve the log odds version of the medical diagnosis task was sufficient to make all of the participants perform almost perfectly. This conclusion could, however, be restricted to these medical diagnosis tasks and not apply to other multiplicative probability problems. Experiment 3 attempted to replicate part of these results with another of the classic probability rules, the computational rule for conjunctive probabilities. This study complements our previous studies on the conjunction fallacy (e.g., Nilsson et al., 2009) by attempting to control the magnitude of the bias with different assessment formats.

With *log probability* the conjunction rule for independent events is additive,

$$\log(p(A\&B)) = \log(p(A)) + \log(p(B)). \quad (9)$$

Log probabilities run on the interval $[-\infty, 0]$. Two examples of natural anchors back to probability is log probability -1 (probability .1) and log probability -2 (probability .01). Log probabilities can be seen as a measure ranging from values of extreme "unlikelihood" (e.g., -3) to certainty (0). If the judged probability that Linda is a feminist is .9, the log probability is $-.05$. If the judged probability that Linda is a bank teller is .1, the log probability is $-.1$. Making the dubious assumption that these possibilities are independent, the log probability for the conjunction is $-.15$, suggesting that the conjunction is more "unlikely".

In Experiment 3 the participants were provided with the stated probabilities of two events and were required to estimate the probability of the conjunction, given the assumption that the events are independent. Half of the participants assessed the problems in a *probability format*, in regard to which the multiplicative integration in Eq. (3) is appropriate; half assessed the problems in a *log probability format*, according to which the additive integration in Eq. (9) is appropriate. For simplicity, all participants first assessed the problems based on a metric instruction, and

then received them a second time with a computational instruction. Note that, in contrast to in most previous studies on conjunction errors (but see also Gavanski & Roskos-Ewoldsen, 1991; Nilsson, 2008), here the participants were provided with explicit numerical estimates of the component probabilities. If conjunction fallacies appear in this task they cannot plausibly be explained by use of the representativeness heuristic.

Because the problems only involve integration of two components and the multiplicative rule (Eq. (3)) for conjunctions of independent events is likely to be known to many participants in an under-graduate student population, we were uncertain as to whether most participants would use controlled intuitive integration, with linear additive integration, or analytic integration based on retrieval from memory of the probability rule (Eq. (3)) and multiplicative facts. Experiment 3 thus corresponds to the comparison between odds (multiplication of two digits) and log odds (addition of two digits) in the previous two experiments, for which the framework in Fig. 1 provides less guidance. Note also that the log odds format will only eliminate conjunction fallacies if people sum the components, not if they spontaneously average them. Therefore, we concentrate on only one of the predictions addressed in the previous two experiments; because the log probability format invites normative integration both with the intuitive default mode and by explicit number crunching, people should find it easier to improve judgment performance with this format.

6.1. Methods

6.1.1. Participants

Thirty undergraduate students participated (17 female and 13 male; average age = 24.3). As compensation, participants received either course credits or a movie ticket.

6.1.2. Apparatus and materials

Stimuli and instructions were included in booklets. The stimuli were 20 scenarios with the following structure:

The log probability that a patient has virus A1 is −0.07.
The log probability that a patient has virus B1 is −0.50.

Each scenario presented the probability that a randomly selected patient has virus A*N* (*N* equaled 1 for the booklet's first scenario, *N* equaled 2 for the booklet's second scenario, and so on) and the probability that a randomly selected patient has virus B*N*. For each scenario, participants were explicitly informed that presence of virus A was completely uncorrelated with presence of virus B. There were two scenario types, probability scenarios (stating each probability as both a number between 0 and 1 and as a percentage) and log scenarios (stating each probability as a log probability; as in the example above). For the probability scenarios the task was to assess the probability in percent that a randomly selected patient has both virus A*N* and virus B*N*. For the log probability scenarios the task was to assess the probability in log probability that a randomly selected patient has both virus A*N* and virus B*N*.

There were three types of instructions, a metric instruction, a computational probability instruction and a computational log probability instruction. The metric instruction included a short text explaining that probabilities can be expressed in various ways. Among other things, the text stated that "probabilities are normally expressed as numbers between 0 and 1 or as corresponding percentages between 0% and 100%" but "are sometimes better expressed as log probability". The metric instructions also included an explanation of the terms logarithm and log probability as well as a table describing how the probabilities of 1%, 10%, 20%, 30%, ..., 90%, 100% can be converted into log probabilities. Thus, in contrast to in Experiments 1 and 2, in Experiment 3 the metric instruction was the same in both the metric probability and the metric log probability conditions, introducing both of the two possible formats.

At the end, the metric instructions included five multiple choice questions concerning the relationship between log probabilities and probabilities described as numbers between 0 and 1 (e.g., the log probability of −1.00 equals which of the following probabilities: .01, .1, .2, .5, 1.0) and a description of the structure of the scenarios. By displaying equations describing the normative combination rules and numerical examples, the computational probability instruction showed how $p(A)$ and $p(B)$ are combined into $p(A\&B)$ and the computational log probability instruction showed how $\log(p(A))$ and $\log(p(B))$ are combined into $\log(p(A\&B))$. There were four types of booklets. The metric probability (log probability) booklet included the metric instruction plus 20% (log probability) scenarios. The computational percentage (log probability) booklet included the computational percentage (computational log probability) instruction plus 20% (log probability) scenarios.

6.1.3. Design and procedure

The experiment was a 2∗2 mixed design with instructions (metric vs. computational; within subjects) and format (probability vs. log probability; between subject) as independent variables. All participants were tested individually. At arrival, participants were handed a booklet. Half of the participants were handed the metric probability booklet and the other half were handed the metric log probability booklet. After completion of the 20 scenarios, participants previously handed the metric probability booklet were asked to complete the computational probability booklet and the other half of the participants were asked to complete the computational log probability booklet.

6.2. Results and discussion

As in Experiment 1, all data was transformed to the 0–1 probability scale. The performance in terms of Mean Absolute Error (MAE) of judgment between the probability assessment and the correct conjunctive probability is summarized in Fig. 7. In the Metric instruction condition, the median MAE is lower (better) with the log probability format (.018 vs. .035 with the probability format), but the difference is not statistically significant (Mann–Whitney; $U = 81$, $Z = 1.03$, $p = .305$). In the Computational instruction condition, however, the median MAE is significantly lower

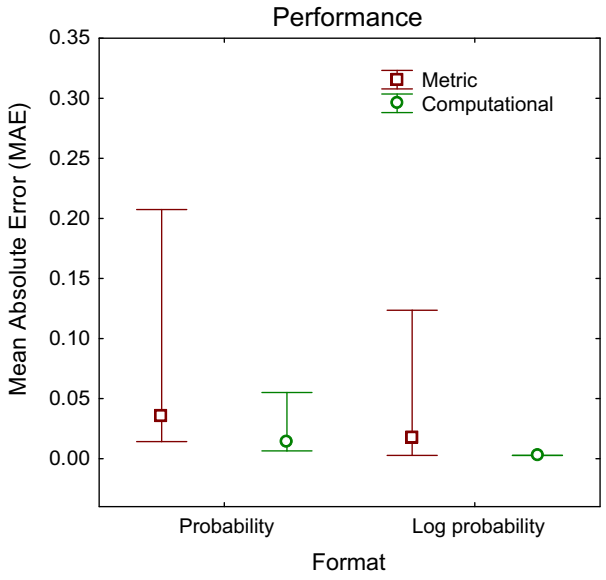

Fig. 7. Experiment 3: The median performance with interquartile index in terms of Mean Absolute Error (MAE) between the judgment and the correct conjunctive probability.

(better) with the log probability format, .003 vs. .014 with the probability format: Mann–Whitney; $U = 27$, $Z = 3.38$, $p = .001$). Again, the performance with the log format and a brief instruction (Metric condition) is virtually as good as performance with the probability format and computational instruction (.018 vs. .014), and performance with log probability and computational instruction is almost perfect.

The mean rate of conjunction fallacies with Metric instructions was 18% with the probability and 25% with the log probability formats (Mann–Whitney; $U = 103.5$, $Z = .04$, $p = .963$). This is lower than in most studies that have not provided the participants with explicit numbers for the component probabilities (where the rate is often in the 30–50% interval, see Nilsson et al., 2009). We attribute this difference to the fact that many of participants in this experiment are likely to have analytic knowledge of the multiplicative integration rule for independent events (Eq. (3)) and the use of explicit numbers probably elicited retrieval of this rule in many of the participants. Notably, the conjunction fallacies for this set of arbitrarily combined statements cannot be attributed to the representiveness heuristic, but is likely to derive from intuitive linear additive integration. Also as predicted, the only conditions where the conjunction fallacies are virtually eliminated is with log probability and computational instruction, where the rate is 1.4% (vs. 6.7% with computational probability format; Mann–Whitney; $U = 57.7$, $Z = 2.05$, $p = .04$). Thus, even with computational instructions, explicitly informing about the correct conjunctive rule from probability theory and providing concrete computational examples, performance with the probability format was significantly poorer than with the corresponding instructions for the log probability format. As in Experiment 1, a log format that made the integration additive served to nominally eliminate the bias.

7. General discussion

The common wisdom in research on probability judgment (Gilovich et al., 2002; Kahneman et al., 1982) is that people often make poor judgments because they rely on simplifying heuristics that substitute variables that are conveniently available (e.g., similarity, fluency) for the normative properties of probability or frequency (Kahneman & Frederick, 2002). An alternative proposal pursued in this article is to place the explanatory emphasis not primarily on the use of variable substitution as such, but on the hypothesis that probability theory is not framed in a way that makes it digestible to the human mind (Juslin et al., 2009). While many probability rules require multiplicative (configural) integration, people certainly seem much more *inclined to*—and with intuitive judgment possibly cognitively *constrained to*—integrate information by linear additive integration (Juslin et al., 2008). The hypothesis as such is not new, but, if anything, one of the best supported conclusions of 50 years of neo-Brunswikian research on judgment (Brehmer, 1994; Cooksey, 1996; Hammond & Stewart, 2001). But its relevance for probability reasoning is often neglected.

The results from the experiments reported in this article generally support the hypothesis that alternative formats that translate originally multiplicative probability problems into formats that require addition immediately decrease the nominal rate of bias. With the term "nominal" we want to emphasize that they are regarded as the predicted effects if people spontaneously integrate information linearly and additively, but they *need not* imply a deep conceptual understanding of the log metrics or an ability to generalize to novel tasks. We do not claim that the short instructions we have provided in our experiments are sufficient achieve these refined levels of understanding. Nonetheless, the results raise the possibility that peoples' judgments could in certain circumstances (e.g., aggregation of risk) be trained to benefit from a format that alleviates the need for multiplication. The degree to which people can, in a deeper sense, learn to think about uncertainty in terms of other metrics than the traditional probability metric is an interesting possibility for future research.

When discussing Experiments 1 and 2 on base-rate neglect it is useful to separate the findings in the probability condition from the comparison between the odds and log odds conditions. The hypothesis that base-rate neglect in the medical diagnosis task is, partially or wholly, explained by the participants using a linear additive integration rule was directly supported by two findings from the probability conditions. First, independently of whether the participants had received an instruction on Bayes' theorem or not, and independently of whether they had received feedback on previous judgments or not, the fit of the linear additive model was better than the fit of the multiplicative model. Second, although both formal instruction and feedback improved performance, it never reached beyond the maximum level allowed by linear additive integration. When participants improve their judgments, either as a result of instruction (Experiment 1) or extensive training

with outcome feedback (Experiment 2), they appear to do so, not by shifting to Bayesian integration, but by adapting the linear additive integration so that it better approximates the output from Bayes' theorem.

An important assumption of the hypothesis developed in this paper is that linear additive integration is the intuitive default that people use when they either lack access to or are unable to implement analytic rules and, therefore, that the demand for multiplicative integration is one factor that often hinders people from implementing probability theory. A memory-free sequential adjustment process can perform accurate linear additive integration of almost any number of cues. However, in a multiplicative task the many interdependencies between the adjustments implied by successive cues rapidly overwhelms the capacities of working memory. In a multiplicative task people are therefore forced to rely on explicit number crunching, or the best linear additive approximation they can muster.

Without any computational instruction it is unlikely that our participants knew what to do when confronted with either the odds version or the log odds version of the medical diagnosis task. Therefore, it is reasonable to assume that these tasks naturally will engage the intuitive default mode of information integration that is applied when no analytic principles are available. If this capacity-constrained and sequential process, as we have argued, naturally implements linear additive integration, better performance will (by mechanism, not design) be achieved with the log odds version. This is exactly what was observed in the experiments.

Further, if multiplicative integration is one factor that obstructs people from implementing normative rules, by necessitating too excessive retrieval and elaboration of declarative facts about multiplication within the constraints of working memory, instructions and feedback should be less effective in a task that demands multiplicative integration than in a task that demands additive integration. This prediction was directly supported by the finding that instructions and feedback had a stronger effect in the log odds conditions than in the odds conditions (remember, the only difference between the odds format and the log odds format is that while the former demands multiplication the latter demands addition).

Both in the odds conditions of Experiments 1 and 2 and in the probability condition of Experiment 3 many participants were able to multiply two digits. These findings parallel those by Nilsson, Rieskamp, and Jenny (2010). Nilsson et al. asked participants to combine constituent probabilities into conjunctive probabilities. They found that while more than a third of the participants multiplied the constituent probabilities, an equally large proportion of participants did so by linear additive integration. Some people thus appear able to multiply the digits, while others have to rely on the default of linear additive integration. The framework in Fig. 1 suggests that a crucial difference between these two groups may refer to their working memory capacity or their motivation. On this view, only persons with a lot of working memory capacity and (or) high motivation are likely to engage in the effortful analytic processes required for explicit number crunching. The data in the current experiments do not allow us to make more detailed tests of the role of working memory and motivation, or to chart the definitive limitations of peoples' ability to perform on line integration according to multiplicative rules. These predictions have to be addressed in future studies.

In addition, the results from Experiments 1 and 2 consistently indicated that spontaneously people typically approach the base-rate problem by integrating the hit-rate and the base-rate, where most weight is placed on the hit-rate, but, surprisingly, on average no weight is assigned to the false-alarm rate. That people assign more weight to the hit-rate is in the spirit of the representativeness heuristic, but in general they also appreciate the importance of the base-rate (Ajzen, 1977; Birnbaum & Mellers, 1983; Fischhoff et al., 1979; Gigerenzer et al., 1988; Kahneman et al., 1982). Reliance on the representativeness heuristic or on a Fisherian algorithm (Gigerenzer & Hoffrage, 1995), as such, provides no rationale for why people should assign *any* weight to the base-rate, which is irrelevant both to the representativeness and the likelihood of the evidence.

Why do people use base-rate and hit-rate but ignore false-alarm-rates? We see two possible (and not necessarily mutually exclusive) explanations. The first is that people have limited ability to consider the implications of the alternative hypothesis in hypothesis testing (Ofir, 1988). The second is that people might consider a failure to diagnose a sick person is more costly than diagnosing a person that is not sick. A second test will correct a false-alarm. The severity of a miss as compared to a false-alarm might be why people overweight hit-rates and ignores false-alarm-rates. If this is the case, it should be possible to change the weighting by changing the framing of the task. That is, if the task is framed such that misses are less costly than false-alarms, more weight might be put on false-alarm-rates than on hit-rates.

The results in regard to the conjunction fallacy in Experiment 3 were similar to the corresponding results in Experiments 1 and 2. Whereas with the additive log probability format, the performance was easily perfected already with a short instruction, the participants struggled to improve their performance with the multiplicative probability format. This parallels the comparison between odds and log odds in Experiments 1 and 2. Importantly, the substantial rates of conjunction fallacies in the Metric condition (22%) cannot be accounted for by the representativeness heuristic, but suggest linear additive integration.

Obviously, there are also limitations of this study. Because of the relatively few observations at the level of individual participants in Experiment 1 there is an obvious risk for overfit in the modeling of the data. To address this problem, we kept the number of free parameters the same in two models that were fitted to data from the condition with probability format and we also applied parameter-free versions of the additive and the multiplicative models (see Footnote 9). We, moreover, note that all important conclusions from Experiment 1 were replicated in Experiment 2, where there were much more observations per participant and the risk of overfit is less of a problem. A second potential limitation that might be raised is that the number of participants in each experiment is rather

low. We, however, note that the main results are, if anything, remarkably consistent across all three experiments.

Altogether we submit that these results suggest that the main constraint on peoples' ability to make the normative judgments (according to probability theory) may not primarily be their reliance on judgment heuristics like representativeness (Kahneman & Frederick, 2002), but their spontaneous inclination for linear and additive information integration. As might be expected if this inclination derives from a cognitive constraint, at least in the more complex problems, like those involving Bayes' theorem in probability version where people are unable to "number crunch" the probabilities according to the analytical formulae, this linear additive disposition appears very little affected by instruction or training.

In this article, the performance was improved by framing the tasks so that linear additive integration could be used to generate normative solutions. It is unlikely that this improvement had anything to do with participants being able to represent the problem in a more correct way. In this respect, our manipulation is conceptually different from most other manipulations that have been found to improve performance in Bayesian reasoning tasks. We believe that this shows that performance in these types of tasks can be improved in two different ways. The first is to frame the task to make it easier to correctly represent the problem, which, in effect, facilitates the usage of analytical processes. The second way is to frame the task in such a way that the information can be combined by linear additive integration (what we suggests is the default rule of the intuitive system).

Performance can be enhanced by manipulations on both the contextual and the computational level. In this way, our findings complement the nested set hypothesis of Barbey and Sloman (2007). In the terminology of the nested set hypothesis, our manipulations aided the processes of intuitive judgment, while manipulations at the contextual level (e.g., natural frequencies) enable participants to apply appropriate analytical processes. More generally, a complete account should explain why these biases occur both when representativeness can and cannot be applied, why the biases are reduced by formats that highlight the set relations, and why the biases can be more easily reduced by additive logarithm formats. We believe that such an account will incorporate both insights captured by the nested set hypothesis (Barbey & Sloman, 2007) and by the computational considerations addressed in this article.

Although our explanation may depart from the typical textbook account, it is less surprising from the perspective of other phenomena in judgment research, such that people successively average "old" and "new" data in Bayesian belief revision tasks (Hogarth & Einhorn, 1992; Lopes, 1985, 1987; Shanteau, 1970, 1972, 1975) and adapt their use of base-rates flexibly depending on context and causal models (Ajzen, 1977; Birnbaum & Mellers, 1983; Fischhoff et al., 1979; Gigerenzer et al., 1988; Kahneman et al., 1982). As noted, a large literature on multiple-cue judgment likewise demonstrates that people are more inclined for linear additive integration than configural integration of cues (Brehmer, 1994; Cooksey, 1996; Hammond & Stewart, 2001). From this point of view, people have problems with many of the classic "cognitive illusions" for the very same reasons as they have problems with many other configural multiple-cue tasks. When presented with judgment tasks for which no analytic principles are available people resort to the cognitive activity of last resort, which apparently fails harmonize with probability theory.

Acknowledgements

This research was sponsored by the Swedish Research Council and the Swedish Tercentary Bank foundation. The authors are indebted to Ebba Elwin, Göran Hansson, and Maria Henriksson for valuable comments and discussions of the topics addressed in this article.

References

Adelman, L., Stewart, T. R., & Hammond, K. R. (1975). A case history of the application of social judgment theory to policy formation. *Policy Sciences, 6*, 137–159.

Ajzen, I. (1977). Intuitive theories of events and effects of base-rate information on prediction. *Journal of Personality and Social Psychology, 35*, 303–314.

Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.

Anderson, N. H. (1996). *A functional theory of cognition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences, 30*, 241–254.

Birnbaum, M. H., & Mellers, B. A. (1983). Bayesian inference: Combining base-rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology, 45*, 792–804.

Brehmer, B. (1994). The psychology of linear judgment models. *Acta Psychologica, 87*, 137–154.

Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago: University of Chicago Press.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications*. San Diego: Academic Press, Inc.

Costello, F. J. (2009). How probability theory explains the conjunction fallacy. *Journal of Behavioral Decision Making, 22*, 213–234.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87–185.

Darlow, A. L., & Sloman, S. A. (2010). Two systems of reasoning: Architecture and relation to emotion. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*, 382–392.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 968–986.

Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review, 112*, 951–978.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judement under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge, UK: Cambridge Univ. Press.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255–278.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational Behavior and Human Performance, 23*, 339–359.

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

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base-rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 513–525.

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

Gigerenzer, G., & Murray, D.J. (1987). Cognition as intuitive statistics. Hillsdale, NJ: Erlbaum. Available from <http://www.kli.ac.at/theorylab/AuthPage/G/GigerenzerG.html>.

Gilovich, T., Griffin, D. W., & Kahneman, D. (2002). *Inferences, heuristics, and biases: New directions in judgment under uncertainty*. New York: Cambridge University Press.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducibly uncertainty, inevitable error, unavoidable injustice*. New York: Oxford University Press.

Hammond, K. R., & Stewart, T. R. (Eds.). (2001). *The essential Brunswik: Beginnings, explications, applications*. Oxford, England: Oxford University Press.

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

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*, 1–55.

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annual Review of Psychology, 59*, 193–224.

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple-cue judgment: A division-of-labor hypothesis. *Cognition, 106*, 259–298.

Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review, 116*, 856–874.

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in multiple-cue judgment. *Journal of Experimental Psychology: General, 132*, 133–156.

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

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgments under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge Univ. Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–251.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134*, 404–426.

Koehler, J. J. (1996). The base-rate fallacy reconsidered: Descriptive, normative and methodological challenges. *Behavioral and Brain Sciences, 19*, 1–17.

Lopes, L. L. (1985). Averaging rules and adjustment processes in Bayesian inference. *Bulletin of the Psychonomic Society, 23*, 509–512.

Lopes, L. L. (1987). Procedural debiasing. *Acta Psychologica, 64*, 167–185.

McKenzie, C. R. M. (1994). The accuracy of intuitive judgment strategies: Covariation assessment and Bayesian inference. *Cognitive Psychology, 26*, 2009–2239.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238.

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

Nilsson, H., Rieskamp, J., & Jenny, M. (2010). Experience with constituent events has no effect on the overestimation of conjunctive probabilities. Unpublished manuscript.

Nilsson, H., Winman, A., Juslin, P., & Hansson, G. (2009). Linda is not a bearded lady: Configural weighting and adding as the cause of extension errors. *Journal of Experimental Psychology: General, 138*, 517–534.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin and Review, 7*, 375–402.

Ofir, C. (1988). Pseudodiagnosticity in judgment under uncertainty. *Organizational Behavior and Human Decision Processes, 42*, 343–363.

Reyna, V. F., & Mills, B. A. (2007). Converging evidence supports fuzzy-trace theory's nested sets hypothesis (but not the frequency hypothesis). *Behavioral and Brain Sciences, 30*, 278–280.

Roussel, J.-L., Fayol, M., & Barrouillet, P. (2002). Procedural vs. direct retrieval strategies in arithmetic: A comparison between additive and multiplicative problem solving. *European Journal of Cognitive Psychology, 14*, 61–104.

Shanteau, J. C. (1970). An additive model for sequential decision making. *Journal of Experimental Psychology, 85*, 181–191.

Shanteau, J. C. (1972). Descriptive versus normative models of sequential inference judgments. *Experimental Psychology, 93*, 63–68.

Shanteau, J. C. (1975). Averaging versus multiplying combination rules of inference judgment. *Acta Psychologica, 39*, 83–89.

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153–160). Cambridge, UK: Cambridge Univ. Press.

Tversky, A., & Kahneman, D. (1983). Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 91*, 293–315.

Villejoubert, G., & Mandel, R. R. (2002). The inverse fallacy: An account of deviations from Bayes' theorem and the additivity principle. *Memory & Cognition, 30*, 171–178.

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.

Wolfe, C. R., & Reyna, V. F. (2010). Semantic coherence and fallacies in estimating joint probabilities. *Journal of Behavioral Decision Making, 23*, 203–223.