Cognition 136 (2015) 282–303

Contents lists available at ScienceDirect

Cognition

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

Is there something special with probabilities? – Insight vs. computational ability in multiple risk combination

Peter Juslin [a,*], Marcus Lindskog [a], Bastian Mayerhofer [a,b]

[a] Department of Psychology, Uppsala University, Uppsala, Sweden
[b] University of Göttingen, Göttingen, Germany

ARTICLE INFO

Article history:
Received 13 May 2014
Revised 27 November 2014
Accepted 30 November 2014
Available online 13 December 2014

Keywords:
Risk integration
Probability reasoning
Multiple-cue judgment

ABSTRACT

While a wealth of evidence suggests that humans tend to rely on additive cue combination to make controlled judgments, many of the normative rules for probability combination require multiplicative combination. In this article, the authors combine the experimental paradigms on probability reasoning and multiple-cue judgment to allow a comparison between formally identical tasks that involve probability vs. other task contents. The purpose was to investigate if people have cognitive algorithms for the combination, specifically, of probability, affording multiplicative combination in the context of probability. Three experiments suggest that, although people show some signs of a qualitative understanding of the combination rules that are specific to probability, in all but the simplest cases they lack the cognitive algorithms needed for multiplication, but instead use a variety of additive heuristics to approximate the normative combination. Although these heuristics are surprisingly accurate, normative combination is not consistently achieved until the problems are framed in an additive way.

1. Introduction

A wealth of evidence suggests that humans are inclined to rely on linear additive combination when making controlled judgments that are constrained by capacity-limited and sequential consideration of cues (Anderson, 1981, 1996; Hogarth & Einhorn, 1992; Juslin, Karlsson, & Olsson, 2008; Lopes, 1985, 1987; Roussel, Fayol, & Barrouillet, 2002; Shanteau, 1970, 1972, 1975). Data on multiple-cue judgment thus typically suggest that judgment is a linear additive combination of the cues (Brehmer, 1994; Cooksey, 1996; Hammond, 1996; Hammond & Stewart, 2001; Juslin, Olsson, & Olsson, 2003; Karelaia & Hogarth, 2008).

This inclination for linear additive combination stands in stark contrast to the requirements for multiplicative combination implied by many of the rules of probability theory. In recent publications (e.g., Juslin, Nilsson, & Winman, 2009; Nilsson, Winman, Juslin, & Hansson, 2009) it has therefore been argued that this propensity for linear additive combination may be an important—and often neglected—constraint on people's ability to reason with probability (see also Jenny, Rieskamp, & Nilsson, 2014; Nilsson, Rieskamp, & Jenny, 2013). Indeed, even classic judgment biases, like the conjunction fallacy and base-rate neglect (Kahneman & Frederick, 2002), may not primarily be explained by use of a specific heuristic per se, like "representativeness", as typically claimed (although people no doubt sometimes use similarity or representativeness to make these judgments), but by a tendency to combine constituent probabilities by linear additive combination. Accordingly, the rate of conjunction errors appears equally high regardless of whether the representativeness heuristic is applicable or not (Gavanski & Roskos-Ewoldsen, 1991; Nilsson, 2008).

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

http://dx.doi.org/10.1016/j.cognition.2014.11.041
0010-0277/

Somewhat surprisingly, perhaps, previous research is not conclusive in regard to whether people have spontaneous appreciation for the reasoning rules specific to probability or if they treat tasks with probability contents just as any other task. One possibility is that extensive experience with the processing of an uncertain environment (in the species or the individual) has "geared" into the mind the rules of reasoning that are specific to uncertainty. The other possibility is that people spontaneously fail to make this distinction and instead apply whatever general-purpose algorithms for reasoning they possess. The issue is further highlighted by the increasing tension (e.g., Bowers & Davis, 2012; Griffiths, Chater, Norris, & Pouget, 2012) between research suggesting that people often violate probability theory (e.g., Gilovich, Griffin, & Kahneman, 2002) and the mounting popularity of Bayesian models of cognition, often presuming that information is processed according to probability theory (Oaksford & Chater, 2006). Although some of this tension may be explained by the research programs being concerned with different levels of analysis (i.e., Griffiths et al., 2012; Marr, 1982), the relationships between these conclusions need to be better elaborated.

There are some indications that information processing that involves probability sometimes departs from processing of other contents. Research in the context of Norman Anderson's functional measurement suggests that, in contrast to the combination in many other tasks, in decision making the combination of probability and value is multiplicative, as implied by expected value and expected utility models (Anderson, 1996). In regard to the decrease in judgment bias often observed when probability problems are presented in natural frequencies, it has been proposed that the cognitive algorithms shaped by evolution are tuned to an input, not in terms of single event probabilities (proportions), but natural frequencies (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995; but see Barbey & Sloman, 2007, for a review and a critical discussion of this argument). It has moreover been suggested that frequencies, clearly a key input to reasoning about probabilities, are encoded and stored automatically (Hasher & Zacks, 1979; Zacks & Hasher, 2002). More generally, it seems fair to conclude that direct comparisons of people's information processing abilities in probability tasks as compared to other formally identical tasks has received very little attention.

In this article, we combine the traditional paradigms from research on probability judgment and multiple-cue judgment in order to investigate whether the observed constraints, or inclinations, in regard to people's combination of cues in multiple-cue judgment extend also to the combination of probabilities, or if there are cognitive algorithms specifically tuned to the input, and the computational demands, relevant to probability. In particular, we will explore if the inclination for linear additive combination applies equally to the processing of probability as it does to standard multiple-cue judgment tasks. In three experiments, we therefore explore people's ability to combine probabilities and compare these abilities to those reported in multiple-cue judgment tasks. In a nutshell: how "special" are probabilities?

1.1. Cue combination in multiple-cue and probability judgment

Multiple-cue judgment is a task where several, often probabilistic, *cues* are used for judgment of some unknown *criterion*, as when a physician relies on symptoms (cues) to make a judgment of the correct diagnosis (criterion). Although multiple-cue judgment tasks may sometimes involve nominal cues (Castellan & Edgell, 1973), making them formally similar to categorization tasks (Juslin et al., 2003), most of the research, and in particular, the research relevant to our current concern, involves the use of continuous ("metric") cues to make a judgment of a continuous ("metric") criterion (Brehmer, 1994). Research on both expert and novice judgment supports a number of fairly robust conclusions (Brehmer, 1994; Cooksey, 1996). First, people often use only a few of the available cues in an inconsistent manner. The same cue values thus elicit different judgments from trial to trial. Second; people often have poor insight into which cues they use and how they combine them (see Lagnado, Newell, Kahan, & Shanks, 2006, for a different view). Third, and as we have seen, the judgments tend to be linear and additive combinations of the cues, and this tendency is pervasive even if the underlying cue-criterion relations violate linearity and additivity (see Karelaia & Hogarth, 2008 for a review).

Research on probability judgment has primarily emphasized the assessment rather than the combination of probability. This is true also of tasks that, at least on the face of it, involve combination of probabilities like the medical diagnosis task (Eddy, 1982):

> 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 with a positive test result has the disease?

Typically, the assessed probability is much higher than the probability implied by Bayes' theorem (here .20), commonly interpreted as the result of a too strong captivation by the diagnostic evidence at the neglect of the low base-rate of the disease (Eddy, 1982; Koehler, 1996). The most influential explanation (Kahneman & Frederick, 2002) emphasizes that people substitute hard "extensional" facts that are relevant to probability, such as frequencies and set relations, with conveniently available, subjective heuristics, like representativeness, which do not obey probability theory (Koehler, 1996; Tversky & Kahneman, 1983).

This medical diagnosis task, which clearly involves the combination of three "cues" (base-rate, hit-rate, and false-alarm rate), is similar to a typical multiple-cue judgment task. It has accordingly been proposed that the inclination for linear additive combination observed in research on multiple-cue judgment might be an important—and often neglected—contributor to biases observed in tasks that require multiplicative probability combination, like the medical diagnosis task (Juslin, Nilsson, Winman, & Lindskog, 2011; Juslin et al., 2009; Nilsson

et al., 2009). In this article, we thus explore the relationship between these two tasks in greater detail.

1.2. Do people understand probability?

In the last 40 years, the popular answer to this question has been—in important respects—no! Apparently, people lack the cognitive algorithms that correspond to probability theory and rely on heuristics that, although useful, produce biases in the probability judgments (Gilovich et al., 2002). An influential view evokes two kinds of processing (e.g., Evans, 2011) where Type 1 processes, which are rapid, intuitive, and heuristic, are supervised by slower, analytic Type 2 processes (see also Evans, 2008; Kahneman & Frederick, 2002). Although Type 1 processing has traditionally been the main seat of heuristic processing, while Type 2 may include analytical insights about probability theory, recent formulations (e.g., Evans, 2011) emphasize that cognitive biases may derive from both processes.

Because, as noted by proponents of dual processing theory (Evans, 2011), all functionally distinct units of behavior involve both Type 1 and Type 2 processing, we find it more useful to distinguish between three judgment processes covering the entire arc from cues to criterion (Juslin et al., 2011). *Analytic judgment* involves manipulation in working memory of explicit representations of numbers and equations (e.g., when solving a Base-rate problem by entering multiplication facts through the steps implied by Bayes' theorem). With *exemplar memory* a judgment is produced by directly retrieving a similar case with a known criterion from memory (e.g., a similar base-rate problem with a known posterior probability). Without computational aids (e.g., a computer), analytic judgment is constrained by working memory and only applicable to the simplest problems, while exemplar memory mainly applies to very well-known judgment domains. Therefore, often people have to resort the third judgment process, aptly described as the "*...cognitive activity of last resort*" (K. R. Hammond).

Controlled intuitive judgment involves considering the impact of each cue on the criterion by a controlled, sequential, and capacity-constrained process. Although the cues are explicitly attended, the cue combination is naturally embodying successive adjustment and the linear additive cue combination that arises is implicit and "emergent" in the process. In such a process, multiplicative (configural) cue combination requires that the adjustment made in view of a cue depends not only on this cue, but also on the cues considered previously in the sequence, which requires taxing working memory resources. Because many probability rules, including the rules for conjunctions and base-rate problems, involve multiplication, people are bound to have difficulties with these problems (see Juslin et al., 2009, 2011). From this point of view, we predict that, from long experience with stochastic events, people do have qualitative insight into many of the probability laws (e.g., that the base-rate is relevant to the (posterior) probability in a base-rate problem, or that conjunctions of events tend to have a low probability and disjunctions of events a high probability). However, in all but the simplest cases,[1] they are unable to perform the multiplicative combination and instead they have to resort to, and adapt, additive heuristics to perform risk combination tasks.

1.3. Multiple risk combination

Risk assessments are crucial to many aspects of everyday life. For example, the planning of a trip involves multiple sources of risk that can influence whether one arrives safely at the destination. The airport shuttle may be late, something may happen on the shuttle ride, or a storm may make it impossible for the flight to depart, and so on. A large number of events may occur and, even if every single event by itself is very unlikely, one has to take into account the accumulation of all of these small risks. People find this very difficult.

When Bar-Hillel (1973) investigated probability judgments for disjunctions and conjunctions of compound events, for example, she found that people tend to overestimate the conjunctive probability but underestimate the disjunctive probability. In the context of linear additive models it is of interest to note that this is predicted if people combine probabilities by averaging, rather than by the rules from probability theory (see also Brockner, Paruchuri, Idson, & Tory Higgins, 2002). Svenson (1984), Shaklee and Fischhoff (1990) and Doyle (1997) investigated risk for compound events in the context of cumulative risk over time. Also here the results were indicative of additive strategies. The data reported in Doyle (1997) are illustrative. He categorized the strategies by combining a statistical analysis of the responses with self-reports from the participants. A lot of participants summed the single year risks. Some participants responded with the average risk for every time period or used the "anchoring and adjustment" heuristic (Tversky & Kahneman, 1974) with single year probability as an anchor and then adjusting it in the direction of the normative response. A negligible minority of the participants applied the normatively correct strategy, and some tried to apply it but failed. In sum, the literature suggests that people have considerable difficulty with multiple risk combination and, if anything, the results are indicative of the use of a plethora of linear additive strategies, like summation, mean, and anchoring-and-adjustment.

1.4. A generic multiple risk combination task

Fig. 1A and B illustrates the task in the experiments reported below, which combine properties of both a multiple-cue judgment task and a probability reasoning task. Participants are asked to observe a sequential production line with four independent steps, where each step is associated with a known probability of error in the production. They are also told that the probabilities of error at each

[1] The simple cases referred to here are those when people can produce the normative answer by use of knowledge retrieved from long term memory, for example, by retrieving the analytic rule that the probabilities of independent events should be multiplied, together with multiplicative facts (e.g., ".5 × . 5 = .25"). While this analytic process might be feasible for some tasks that require conjunctive combination of a few probabilities, they are less plausible in tasks with many error sources or a disjunctive frame (see Juslin et al., 2011).

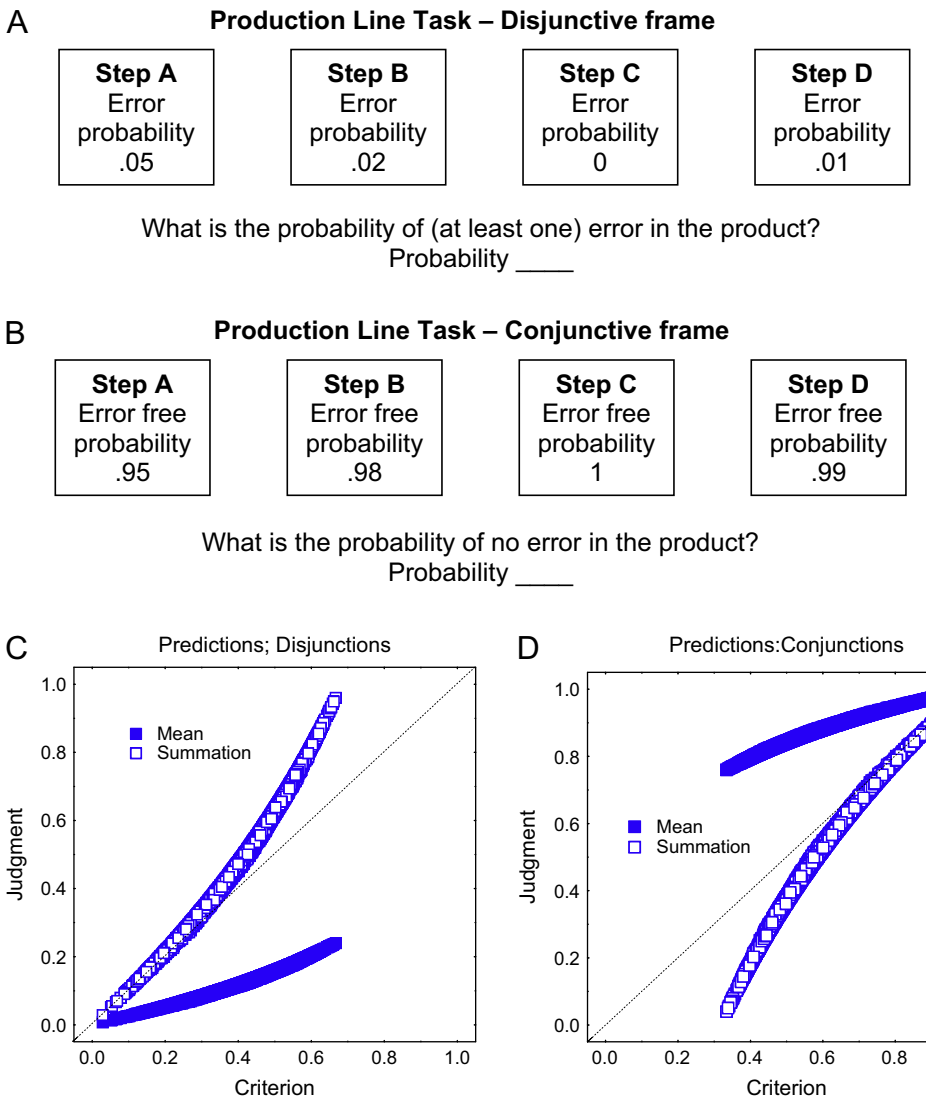

Fig. 1. Panel A: Schematic illustration of the multiple-risk combination task in disjunctive frame used in Experiment 1. In a production line there are a number of production steps, where there is an independent probability (risk) of an error occurring in each production step. The participant's task is to assess the probability of (at least one) error in the final product. Panel B: Schematic illustration of the corresponding multiple-risk combination task in conjunctive frame. Here the task presents the probability of error-free production in each step of the process and the task is to assess the probability of error-free production in all four steps. Panel C: The predictions by a summation and a mean heuristic plotted against the normative probability in the multiple-risk combination task in disjunctive frame. Panel D: The predictions by a summation and a mean heuristic plotted against the normative probability in the multiple-risk combination task in a conjunctive frame. The predictions in C and D were computed by sampling 10,000 quadruples of error probabilities, where each error probability was independently sampled from a uniform probability between 0 and .25.

production step are independent; in other words, that an error occurring at one step does not affect the probability of error at another step. The task in the experiments frames the probability information in two different ways. In the *disjunctive frame* the probability of an error is provided at each step and participants are asked to report the probability of at least one error occurring in the final product that has passed through all four steps. Participants are thus required to estimate the *disjunction* of the four error probabilities. In the *conjunctive frame* the probability of no error (i.e. the complement of the probability of an error) is provided at each step and participants are asked to report the probability that no error has occurred in the final product. This frame requires the participants to estimate the *conjunction* of the four probabilities of no error.

The normative solution in both frames requires multiplicative combination of the probabilities. In the disjunctive frame in Fig. 1A, which reports error probabilities, the probability $p(E)$ of (at least one) error in the final product, after passing through all production steps, is a multiplicative function of the error probabilities $p_i(e)$ in each step ($i = 1,\ldots,4$),

$$p(E) = 1 - \prod_{i=1}^{4}(1 - p_i(e)). \quad (1)$$

In the conjunctive frame in Fig. 1B, that reports the probability of error-free production at each step, the probability $p(C)$ of a correct (or error-free) product is computed,

$$p(C) = \prod_{i=1}^{4} p_i(c), \quad (2)$$

where $p_i(c) = 1 - p_i(e)$.

1.5. Additive heuristics for risk combination

In the disjunctive frame, even if the participants lack knowledge of the exact functional forms of probability theory they might possess the qualitative insight that; as the error probabilities in each step increase, the total probability of an error increases. One way to accommodate this insight, without knowing Eq. (1), would be to add up the error probabilities at each step. In the disjunctive frame, corresponding to Eq. (1), this implies,

$$S_d = \sum_{i=1}^{4} p_i(e). \quad (3)$$

For example, when faced with the task illustrated in Fig. 1A, adding the four risks would yield a probability of at least one error of .08 (.05 + .02 + 0 + .01). When applied to small error probabilities summation is an accurate heuristic for computing the total risk.[2]

Fig. 1C illustrates the accuracy of risk summation in a disjunctive frame. As is evident in Fig. 1C, for normative risks below .2, summation and probability theory yield virtually identical results. At higher probabilities summation overestimates the normative risk, although the overall correlation between the sum and the normative risk is .99. Fig. 1C also illustrates a mean heuristic, using the average error probability A_d as the estimate of the total error probability (i.e., $A_d = S_d/4$), which underestimates the normative risk. Note also that it is very straightforward to adapt these heuristics so that they not only correlate .99 with the normative answer, but also coincide numerically with the normative answer. For example, in the range shown in Fig. 1C, learning to report approximately 3/4 of the sum risk probability yields a Mean Absolute Deviation (MAD) from the correct answer lower than .04.

In a conjunctive frame, participants are faced with the probabilities of no errors in each step and are asked to give the probability that no error occurs in the production line. Even though the normative combination of the probabilities is quite straightforward it requires both knowledge of Eq. (2) and the ability to carry out the multiplication. In principle, people could use a corresponding summation heuristic also here. If people have the qualitative insight that the probability of no error in the production is inversely related to the probability of error at each step, they could sum the difference between 1 and the stated probability at each step and reduce a perfect (1) probability of no error with this amount, that is:

$$S_c = 1 - \sum_{i=1}^{4} (1 - p_i(c)). \quad (4)$$

As with summation in the disjunctive frame, this formulation implements the insight that the overall probability of error is a positive function of the errors at each step, but as evident from Eq. (4) the transformations and computations implied in a conjunctive frame are almost as complex as the normative combination and therefore it does not appear plausible as a heuristic.[3] Fig. 1D illustrates the accuracy of the summation heuristic in Eq. (4) and a mean heuristic ($A_c = 1 - 1/4 \cdot \sum_{i=1}^{4}(1 - p_i(c)) = \sum_{1}^{4} p_i(c)$) in a task with conjunctive frame. In the conjunctive frame the mean heuristic is much easier to use than the summation heuristic, because it can be directly applied to the stated probabilities ($p(c)$), whereas the summation heuristic requires taking the complement of the stated probabilities. In sum: in the disjunctive frame it is plausible that people can implement both the summation and the mean heuristic, although the mean heuristic implies a lot of bias in the judgments in absolute terms. In the conjunctive frame, the summation strategy is too complex to be a plausible heuristic, while the mean heuristic is easy to implement but also here of more limited validity.

The heuristics described above have limitations, of course. As noted above, summation works best with small probabilities and under certain circumstances they produce unreasonable results. For example, in the disjunctive frame with four error probabilities, each larger than .25, summation produces a combined probability larger than 1. In practice, for the heuristics to deliver both simple calculation and approximately valid output, we expect the summation heuristic primarily to be applied to small error probabilities in a disjunctive frame and the mean heuristic to be applied primarily in the conjunctive frame.

The results illustrated in Fig. 1C and D suggests two conclusions. A theoretical conclusion is that there may be little incentive to shift from summation to multiplication, or even that it might be difficult to detect the superiority of the latter strategy from experience (Juslin et al., 2009). A methodological conclusion is that in data with measurement error, the strategies will be very difficult to distinguish on the basis of model fits alone.

1.6. Overview of the experiments

In Experiment 1, we compare two formally identical tasks, one that involves risk (see Fig. 1A and B), the other a standard multiple-cue judgment task involving inference about a non-stochastic criterion. Experiment 2 addressed if the multiplicative nature of the risk combination task is easier to detect if we effectively concentrate the predictors

[2] To see why summation is a good heuristic, it is useful to consider an alternative way to represent the disjunctive probability in Eq. (1): as the sum of the probabilities minus their intersection. The equation below computes the probability $p(A \vee B)$ of the disjunction of events A and B,

$$p(A \vee B) = p(A) + p(B) - p(A) \cdot p(B).$$

When $p(A)$ and $p(B)$ are small the intersection is small and a summation strategy comes close to the normative answer. However, as $p(A)$ and $p(B)$ increase the last term also grows larger. Accordingly, with large error probabilities, the intersection is large and the summation strategy overestimates the total risk.

[3] In principle, of course, in the conjunctive frame one could envision also the possibility that the participants sum the probabilities of error-free production rather than the error probabilities. However, at least in this task this strategy produces a very poor heuristic with little rationale, which in general will imply overall probabilities larger than 1. We found no evidence for such strategies in the experimental data that is reported below.

of the risk to fewer risk sources (cues) and if the training phase highlights the regions where multiplicative and additive combination yields the largest discrepancies. In Experiment 3, we explore the possibility to help the participants by turning to a measure of risk that is inherently more compatible with the human propensity for linear additive risk combination.

2. Experiment 1: multiple-cue vs. risk judgment

In standard multiple-cue judgment tasks, participants are inclined to rely on linear additive combination. Experiment 1 was a first attempt to investigate if this propensity extends to a multiple risk combination task, like the one illustrated in Fig. 1A. We directly contrast a formally identical multiple-risk task to a multiple-cue judgment task. In both tasks, the same information about the cues is conveyed in the instructions to participants. As noted above, the plausible strategies are highly correlated and difficult to distinguish. In addition, to computational modeling that fits multiplicative and additive models to the judgments, we therefore also report an analysis of the residuals from the normative combination.

If the participants use the normative combination in Eqs. (1) and (2), the conjunctive frame would seem to be an easier task than the disjunctive frame, because it involves the straightforward multiplication of the probabilities of no error stated in the task (Eq. (2)). But if people use summation, the disjunctive frame seems the easier task, because they only have to sum the stated probabilities, while the conjunctive frame involves the additional step of first computing the complements of the probabilities stated in the task. The mean heuristic is equally easy to apply in both frames, but it allows relatively poor performance.

In the experiment, the participants first receive a pretest without feedback, then a training phase where they make judgments and receive outcome feedback, finally they receive a posttest. Therefore, rather than relying on rule-based combination (additive or multiplicative), exemplar memory (Medin & Schaffer, 1978; Nosofsky & Johansen, 2000) potentially becomes another way to make the risk judgments by retrieving memory traces of similar previously encountered risk configurations together with their correct overall risk. To investigate if the participants rely on exemplar memory at test, they are also required to make judgments of risk probabilities outside of the training range (i.e., to extrapolate). Exemplar models predict that, because the judgments are a weighted average of the risk probabilities observed in training, the participants should be unable to extrapolate outside of the training range. However, if they rely on some abstract rule from probability theory, or some additive heuristic, they should be able to extrapolate beyond the training range (DeLosh, Busemeyer, & McDaniel, 1997; Juslin et al., 2003 for discussion of such tests of exemplar models).

2.1. Method

2.1.1. Participants

Participants were 40 undergraduate (14 male and 26 female) students from Uppsala University (M = 24.4 years, Sd = 4.0). They received a movie voucher or course credits as compensation for participating in the study.

2.1.2. Design

Experiment 1 used a $2 \times 2 \times 2$ split-plot design with Frame (conjunction/disjunction) and Task (multiple-cue/risk) as between-subject variables and Training (pretest/posttest) as within-subjects variable.

2.1.3. Materials and procedure

The computerized task was divided into three parts, pretest, training and posttest, and carried out on a PC. On each trial in all three parts participants were presented with the combination task illustrated in Fig. 1A.

In the risk condition, the participants task was to assess the probability (in percent) of at least one error (disjunctive frame) or the probability of no error (conjunctive frame) in the final product of a whiskey production line, when the independent probability (risk) of an error (disjunctive frame) or the probability of no error (conjunctive frame) occurring in each of four production steps (A–D) was given. In the multiple-cue condition, the participants' task was to assess the concentration in the blood stream (in percent) of fictitious substance α (disjunctive frame) or the percentage that was not fictitious substance α (conjunctive frame), when the proportion of blood cells containing (disjunctive frame) or the proportion of blood cells not containing (conjunctive frame) one of four (A–D) virus strains was given. The function relating the proportion of substance α and the proportion of the four virus strains was given by Eqs. (1) and (2) in the disjunctive and conjunctive frames, respectively. Thus, in both conditions, the participants are presented with information that involves an identical metric. However, in the risk condition the cover story involves assessment of a stochastic component of a probabilistic process where the cues are probabilities; in the multiple-cue condition the cues are proportions that deterministically determine the α-concentration in the blood.[4]

Careful attention was given to ascertain that comparable information was available both in the risk and the multiple-cue judgment task. The cover story in the risk task involving risks of error in a production line naturally conveys that there is a positive relationship between the risk of error in each production step and the total risk of an error in the product. In order to convey comparable prior information in the multiple-cue judgment task the instruction told the participants that it was known that higher concentrations of the four virus strands were associated with higher levels of the substance α. In both the risk and multiple-cue condition participants were told that the cues were independent: there was no relationship between the level of one cue and the level of the other cues. The two tasks were thus identical, except that in

[4] Note that although probability, from a frequentistic perspective, can be viewed as the long-run proportion of an event, a proportion as such does not imply a stochastic component. To introduce uncertainty and a stochastic component, an element of random sampling from a reference class has to be introduced. There need not be anything uncertain about the claim that, for example, the blood stream contains 10% of a substance.

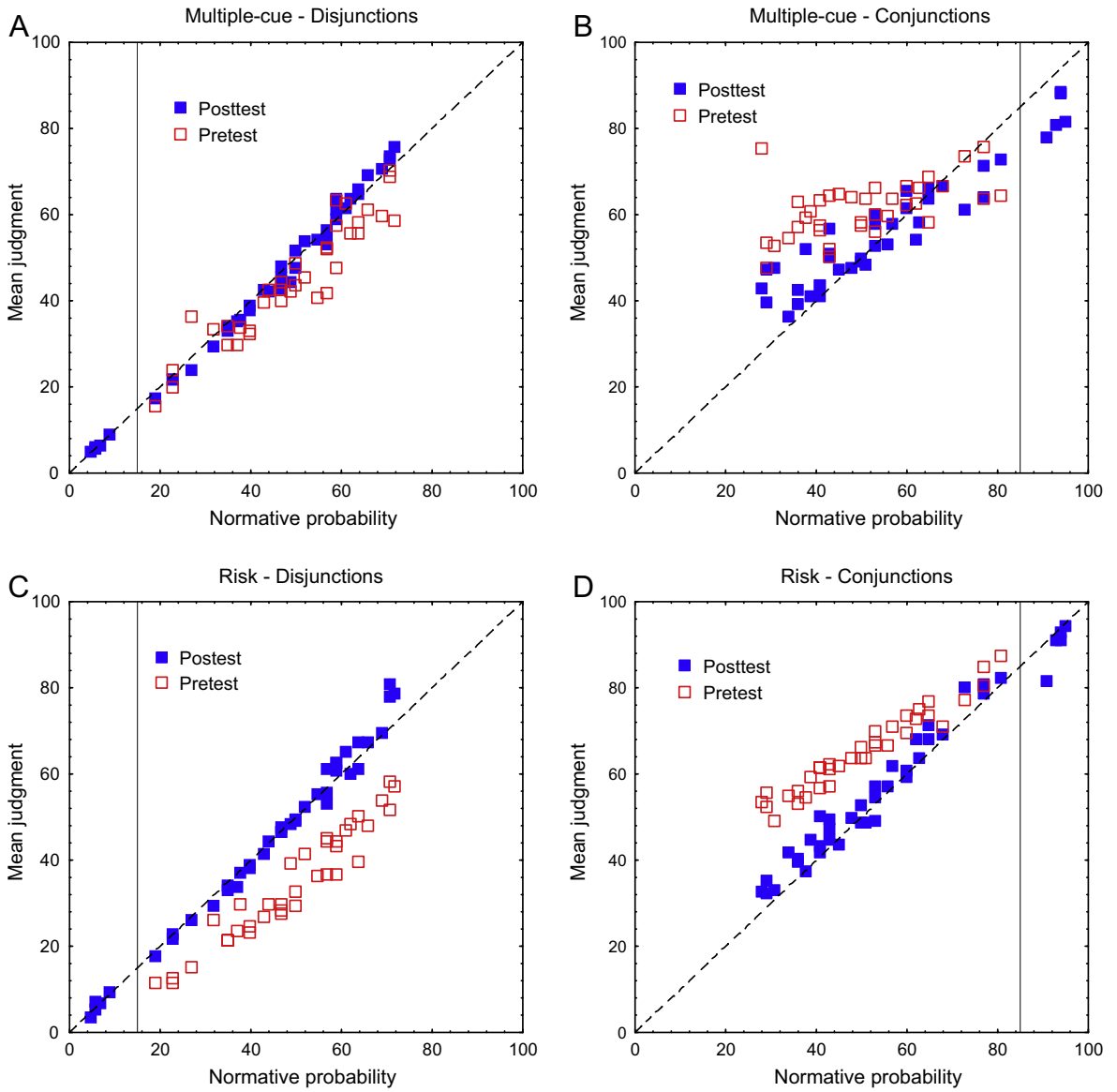

Fig. 2. Experiment 1: Mean judgments for each item in the pretest and the posttest plotted against the normative criterion in the conditions with multiple-cue judgments in a disjunctive frame (Panel A), multiple-cue judgments in a conjunctive frame (Panel B), multiple-risk judgments in a disjunctive frame (Panel C), and multiple-risk judgments in a conjunctive frame (Panel D). In Panels A–D, the identity line represents accurate judgments and the vertical line delineates the extrapolation region (to the left of the line for disjunctions and to the right of the line for conjunctions).

the risk condition the variables were probabilities, while in the multiple-cue task the variables were (deterministic) proportions of substances in the blood stream.

Stimulus for the pre- and post-tests was created by orthogonally combining cue values for each of the four cues (C1: .1, .3, .5; C2: .05, .2, .35; C3: .05, .1; C4: 0, .05) giving a total of 36 items. In the disjunctive frame participants were presented with one of these probabilities/proportions for each cue while participants in the conjunctive frame were presented with their complement (1-C). Using Eqs. (1) and (2) gives criterion values in the intervals [.15, .68] and [.32, .85] in the disjunctive and conjunctive conditions respectively. Two sets of 60 items was created as stimulus for training by randomly drawing error-free probabilities on the range [.8, 1.0] for cue 1 and 2 and [.6, 1.0] for cue 3 and 4 with the constraint that the total error-free probability should be in the range [.35, .85]. Participants were randomly assigned to one of the two training sets. In pretest participants judged 36 items without feedback. In training participants made 60 judgments with feedback of the correct criterion following each judgment. Finally, in posttest participants made judgments on the 36 items from pretest and on five additional items that require extrapolation, below .15 for disjunctions and above .85 for conjunctions, which are diagnostic of exemplar memory. On each trial all four cues were presented

Table 1
Results of a 3-way repeated-measures ANOVA on performance (Root Mean Square Deviation, RMSD) in Experiment 1, where frame and task content are between-subjects variables and training (pretest vs. posttest) is a within-subjects variable with effect size (partial η^2_p). The statistically significant effects are highlighted in bold characters.

	SS	*df*	*MS*	*F*	*p*	η^2_p
Frame	.0116	1	.0116	1.533	.224	.041
Task content	.026	1	.026	3.381	.074	.086
Frame × Task content	**.068**	**1**	**.068**	**8.953**	**.005**	**.199**
Error	.272	36	.008			
Training	**.419**	**1**	**.419**	**69.602**	**.000**	**.659**
Training × Frame	.000	1	.000	0.031	.862	.001
Training × Task content	.003	1	.003	0.516	.477	.0141
Three-way interaction	.004	1	.004	0.666	.420	.0182
Error	.217	36	.006			

simultaneously, as illustrated in Fig. 1A and B, with cues C1–C4 occupying the production steps A–D respectively. The presentation order of items within each part of the experiment was randomized for each participant.

2.2. Results and discussion

Fig. 2A–D plots the mean judgment for each item against the criterion separately for the pretest and posttest in each of the four task-event conditions. The identity line represents accurate judgments and the vertical line delineates the extrapolation region (to the left of the line for disjunctions and to the right of the line for conjunctions). These figures support at least three conclusions. First, in pretest most participants start off with the judgment bias that is consistent with a mean heuristic: underestimation for disjunctions and overestimation for conjunctions. Second, in all conditions except the multiple-cue conjunctions the participants make very accurate judgments in the posttest. Third, in all four conditions the participants extrapolate beyond the training range, accurately estimating the criterion values below .15 with the disjunctions and over .85 for disjunctions, despite that they have never encountered criterion values which are that extreme in training. This suggests that they rely on rule-based combination rather than on exemplar memory. In the following, we first report measures of performance in terms of Root Mean Square Deviation (RMSD) between the judgment and the criterion. Thereafter, we report the results of fitting multiplicative and additive models to the judgments by the individual participants, together with the results of an analysis of the residuals from normative combination.

2.2.1. Performance

RMSD was entered into a split-plot ANOVA with Task (multiple-cue vs. risk; between-subjects), Frame (conjunction vs. disjunction; between-subjects) and Training (pretest vs. posttest; within-subjects) as the independent variables. Table 1 shows that the only statistically significant effects were a main effect of Training ($F(1,36) = 69.60$, $MSE = .006$, $p < .001$, $\eta^2_p = .659$) and an interaction effect between Task and Frame ($F(1,36) = 8.695$, $MSE = .008$, $p = .005$, $\eta^2_p = .199$). As illustrated in Fig. 3, these results document a difference between the multiple-cue and the risk judgments: With multiple-cue judgments, the disjunctive frame leads to better performance than

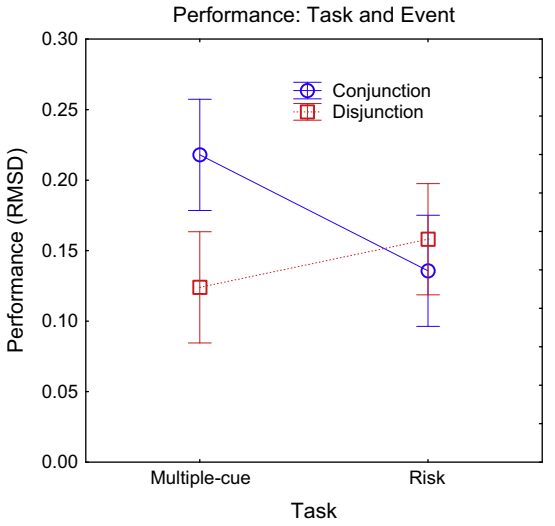

Fig. 3. Mean performance in terms of Root Mean Square Deviation (RMSD) from the criterion with 95% confidence intervals ($N = 10$) in conditions with a multiple-cue or a risk content of the task, separately for conditions with conjunctive and disjunctive framing.

the conjunctive frame, but with risk judgments performance is similar with both frames. The order of difficulty with multiple-cue judgments but not with risk judgments thus coincides with the order predicted by the use of a summation heuristic, because with disjunctions summation can be applied without computing complements of the probabilities explicitly stated in the problems. Performance also improved substantially in all conditions (RMSD decreased from .231 to .087). As noted in connection with Fig. 2A–D, this improvement mainly stems from reduction in initial bias.

2.2.2. Modeling

Two models were individually fitted to the judgments by each participant. The first model is a generalization of the normative model from probability theory,

$$M_{ij} = m \cdot p + (1 - m) \cdot .5, \quad (5)$$

where M_{ij} is the predicted judgment by participant i for task item j, p is the normative response, and m is a free parameter that captures imperfect learning of the

normative response from a uniform prior probability distribution for the value of p. In the disjunctive frame p is obtained from Eq. (1), while in the conjunctive frame p is obtained from Eq. (2). With $m = 1$, the participant performs the normative assessment, whereas $m = 0$ implies that the participants always responds with the expected value of a uniform prior for p in the interval 0–1 (i.e., .5). Eq. (5) is the normative combination, but values of m between 0 and 1 allow also for partial learning and for imperfections in the execution of the normative combination.

The second model implies that the participants perform additive combination of the probabilities/cues stated in the problem,

$$A_{ij} = a \cdot (p_1 + p_2 + p_3 + p_4), \quad (6)$$

where A_{ij} is the predicted judgment by participant i for task item j, p_1, p_2, p_3, and p_4 are the four probabilities/cues stated in the task, respectively (see Fig. 1A and B), and a is a free parameter that captures the mode of additive integration, where $a = 1$ is summation and $a = .25$ is a mean. In the disjunctive frame the probabilities p_1, p_2, p_3, and p_4 are error probabilities $p(e)$ and in the conjunctive frame the probabilities refer to error-free production $p(c)$.[5]

Both of the models were fitted individually and separately for the Pretest and the Posttest. The parameters a and m were left unconstrained and the Levenberg–Marquardt procedure in the Nonlinear Estimation module of the Statistica software package was used to find values that minimized the mean sum square of prediction error (*MSE*). Table A1 in Appendix reports the medians for *MSE* and the parameters (m or a) across participants. The model fit for both models for each individual participant is presented in Fig. 4. Both the measures for model fit and the parameters were characterized by skewed distributions and heterogeneous variance, requiring the use of on nonparametric statistics.

In Fig. 4A and B we see that in Pretest there is large variability in the model fit. There were no statistically significant main effects of Task and Event on the model fit for the multiplicative model, nor for the additive model (Mann–Whitney Tests, all *ps* > .27). The additive model provided better fit (lower *MSE*) than the multiplicative model in all four cells (Wilcoxon Test, $N = 40$, $T = 126$, $Z = 3.817$, $p < .001$, across all four cells).

Fig. 4C and D illustrates that in the Posttest there is less variability and the models provide better and often similar fit (as expected given the high correlation between the predictions). In regard to neither model was there a statistically significant effect of Task on model fit (Mann–Whitney Tests, all *ps* > .473), but the multiplicative model provided significantly better fit in the disjunctive than in the conjunctive frame (Mann–Whitney Test, $N = 40$, $U = 109$, $Z = 2.448$, $p = .014$), as did the additive model (Mann–Whitney Test, $N = 40$, $U = 5$, $Z = 5.261$, $p < .001$). The model fits of the two models were thus compared separately in each cell of the design and the multiplicative model has somewhat better fit in the risk task with a conjunctive frame (Wilcoxon Test, $N = 10$, $T = 2$, $Z = 2.599$, $p = .009$), whereas model fit is similar or identical in the other three cells (Wilcoxon Test, *all ps* > .168). The proportion of participants best predicted by the additive model decreased from 78% in the Pretest to 42% in the Posttest (this decrease is statistically significant, $p = .001$. The Pretest proportion is statistically significant from a H_0 of .5, $p = .002$, the Posttest proportion is not, $p = .08$).

In sum: in the Pretest there were no statistically significant effects of Task or Event on the fit of the models and overall the additive model provided better fit to data. In the Posttest the models provided similar fit, except that there were indications that the multiplicative model provided somewhat better fit in the risk task with a conjunctive frame (i.e., as supported by the pattern in Fig. 4D and the significantly lower *MSE* for the multiplicative model in this condition). In the following, we examine the performance in the Posttest more carefully by illustrating the strategies and performance by individual participants and by analyzing the residuals between the judgments and the normative response.

2.2.3. Examples of individual strategies

Fig. 5 illustrates the performance by individual participants in the multiple-cue judgment task in a disjunctive frame and Fig. 6 illustrates the performance by the individual participants in the risk task in a conjunctive frame (the other two cells are similar, but for space reasons we refrain from producing them here). Fig. 5 illustrates that, although performance in the Posttest with disjunctive frame was impressive, most of the functions preserve the accelerated shape with over-estimation for the high values that is characteristic of additive integration, and especially of summation (e.g., ID 1).

This appears to define a strategy of *proportional summation*, where the participants sum the risks and report some proportion (e.g., 80%) of this sum to minimize the overestimation that would otherwise occur (see predictions for summation in the larger panel of Fig. 5). Two exceptions are ID 20 and ID 30, which rely on a strategy of *truncated summation*, where summation is used up to a ceiling, after which all higher sums are truncated at this level, close to .70. These participants seem to have learned from the training that summation is a good approximation of the criterion, but that the criterion values never go higher than .70.

In Fig. 6, we see that in the risk task with conjunctive frame, at least, some of the participants seem to reproduce the normative values (e.g., ID 9 and ID 34), while other participants still follow an additive strategy in the posttest, with deviations that are characteristic of the mean heuristic (e.g., ID 12 and ID 22). Notably, the risk task with conjunctive frame is the only condition where there are signs of some participants truly approximating the normative combination and indeed, as we have seen, this was the

[5] Note that Eq. (6) takes only into account additive combination of the *stated* probabilities/cues. In principle, some of the participants could be adding up the *complements of the stated probabilities* (as in Eq. (4)). This is a less sensible heuristic given that it amounts to a combination strategy that appears virtually as complex as the normative combination. However, three participants in the Pretest were somewhat better described by such a model summing up the complements to the stated probabilities/cues than by Eq. (6). This was true of no participant in the Posttest of Experiment 1 and never occurred in Experiment 2 below.

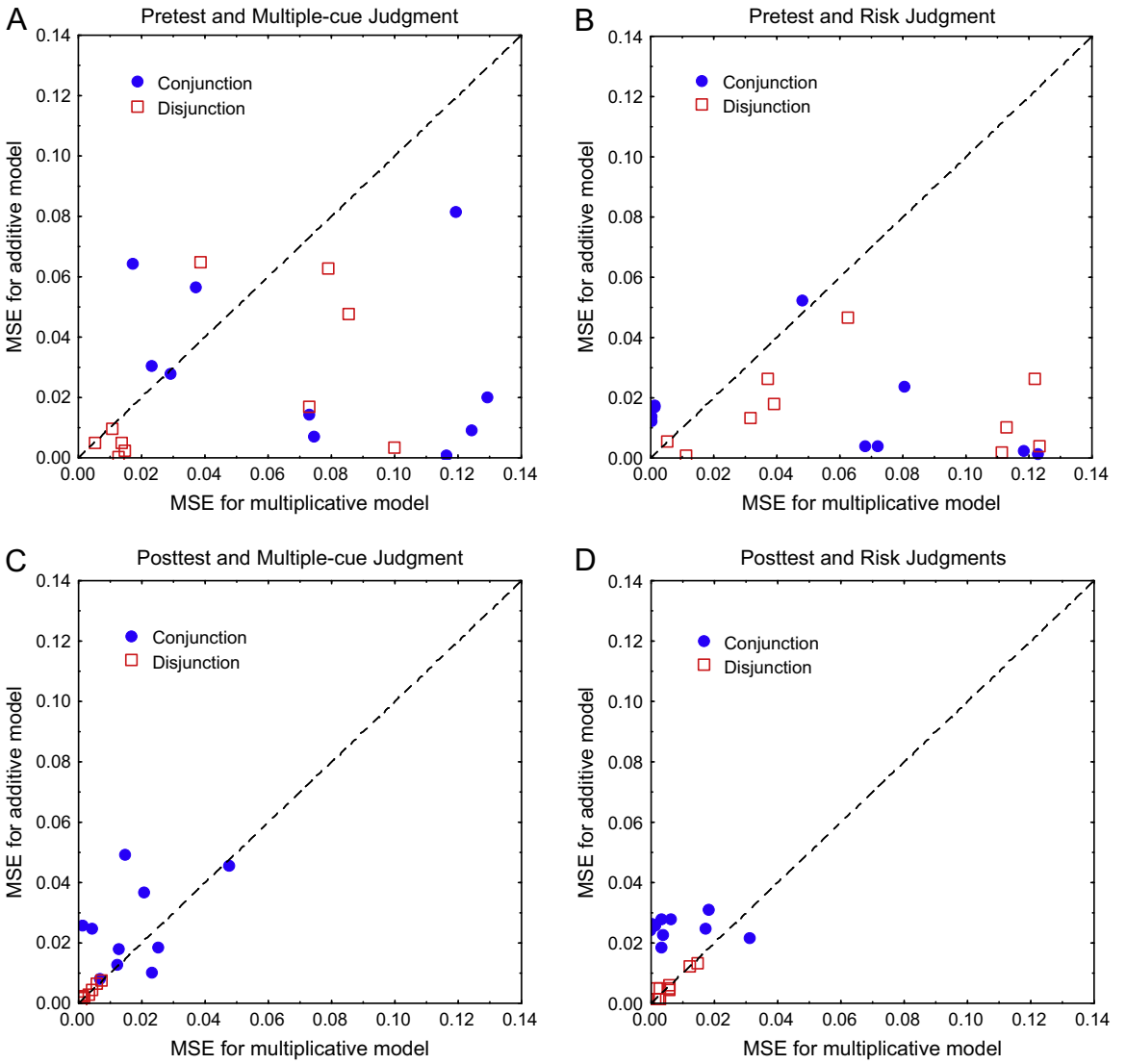

Fig. 4. The model fit in terms of Mean Square Error (*MSE*) between predictions and data for the multiplicative model on the *x*-axis and for the additive model on the *y*-axis, separately for each individual participant in each condition of Experiment 1. Data points above the identity line indicate better fit for the multiplicative model and data points below the identity line better fit for the additive model.

only cell in Experiment 1 where the multiplicative model provided better fit.

2.2.4. Residual analysis

To further analyze the Posttest judgments, we investigated the residuals between judgments and the normative multiplicative model. If participants rely on multiplicative combination, we should expect these residuals to be small and nonsystematic; randomly hovering around 0 at all levels of the normative value. However, as illustrated in Fig. 1C, in the disjunctive frame use of an additive heuristic should produce a nonlinear pattern of residuals, with larger deviations for the higher normative values: viz. overestimation in the case of the summation heuristic and underestimation in the case of the mean heuristic. The residuals predicted if participants use summation are therefore positively correlated with the normative values (more positive residuals, the higher the normative value) and the residuals predicted if they use the mean heuristics are negatively correlated with the normative values (more negative residuals, the higher the normative value). The residuals from use of the summation and the mean heuristics are thus negatively correlated.

Fig. 7 plots the mean residuals between the judgments and the criterion (the normative value) against the criterion across all 41 items that were part of the posttest. In all four posttest conditions, the residuals are not random, but substantially and significantly correlated with the criterion (the correlations are reported in the panels of Fig. 7). With the disjunctive frame, the correlations are positive with an accelerating trend that is suggestive of the summation heuristic. For multiple-cue judgment with

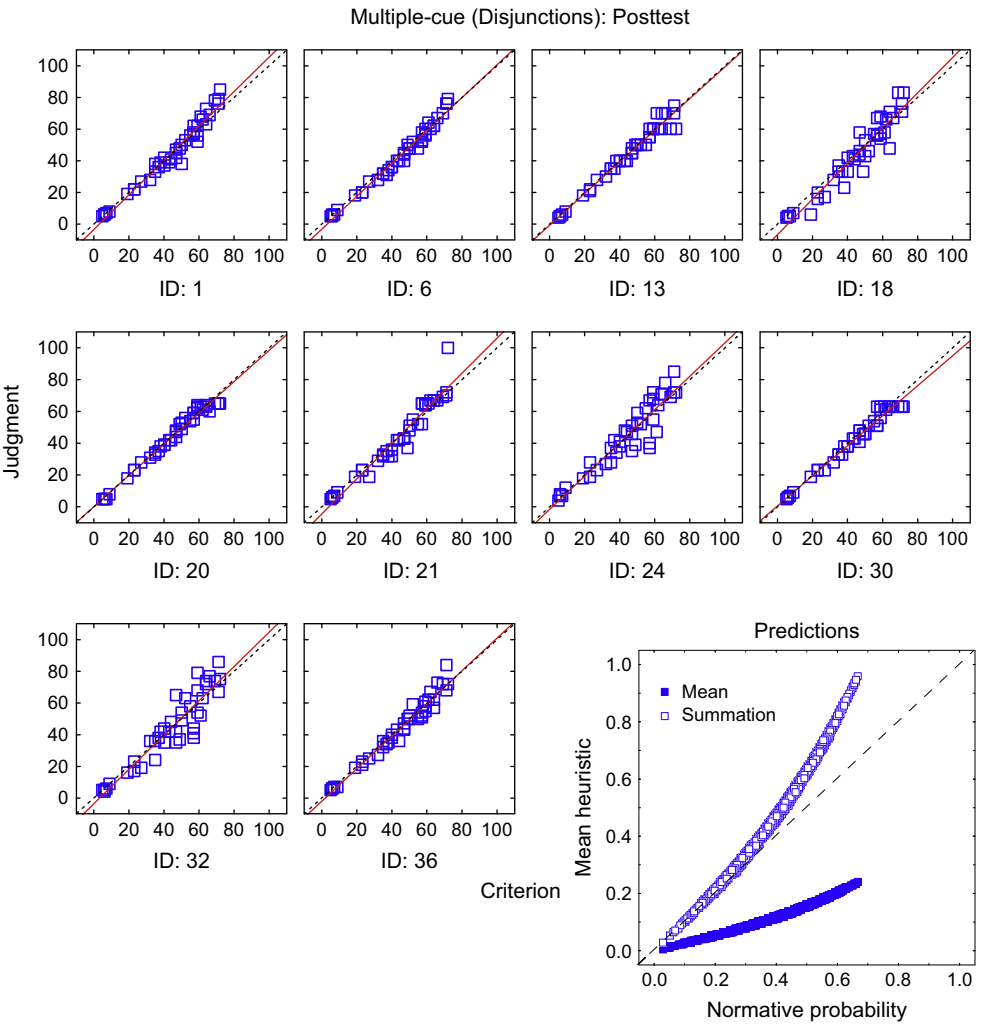

Fig. 5. The judgments by each participant plotted against the normative value in the Posttest with a multiple-cue judgment task and a disjunctive frame, together with the predictions by a mean and a summation heuristic (the larger lower panel).

a disjunctive frame, the correlation between the mean residuals from the normative response and the residuals predicted by the summation heuristic was .603 ($p < .001$) and for the risk judgments with a disjunctive frame this correlation was .704 ($p < .001$). Because these correlations are computed across means for each item, the correlations could, in principle, be consistent with a majority responding in the normative manner, where the correlation is driven by only a few (or even one) participants that use summation. However, the corresponding median correlation is .308 for multiple-cue judgment and .359 for risk judgment across individual participants ($t_{18} = .754$, $p = .461$ for the difference between the multiple-cue and risk conditions)[6]. Across all 20 participants with a disjunctive frame the median correlation was .351 ($t_{19} = 2.997$, $p = .008$, given $H_0 = 0$). The residual analysis in Fig. 7 thus allows us to reject the hypothesis that the improvement with training observed with the disjunctive frame was obtained primarily by a shift to the use of multiplication, but instead suggests the use of some accommodation of summation.

As illustrated in the lower panels of Fig. 7, with a conjunctive frame the residuals are negatively correlated with the criterion, suggestive of the use of a mean heuristic. Across the 41 items in the posttest with a conjunctive frame, for the multiple-cue judgments the correlation between the mean residuals from the normative response and the residuals predicted by the mean heuristic was .829 ($p < .001$) and for the risk judgments this correlation was .468 ($p = .002$). When the correlation is computed separately for each participant in the conjunctive frame, the median correlation is .551 for multiple-cue judgment and .193 for risk judgment ($t_{18} = 1.96$, $p = .066$ for this difference between the conditions). Across all 20 participants with a conjunctive frame, the median correlation was .283 ($t_{19} = 3.221$, $p = .004$, given $H_0 = 0$). The residuals of

[6] The correlations were squared, but with preserved sign for the direction of the relationship, before they were entered into the *t*-test. The same holds for the corresponding analyses of correlations reported below.

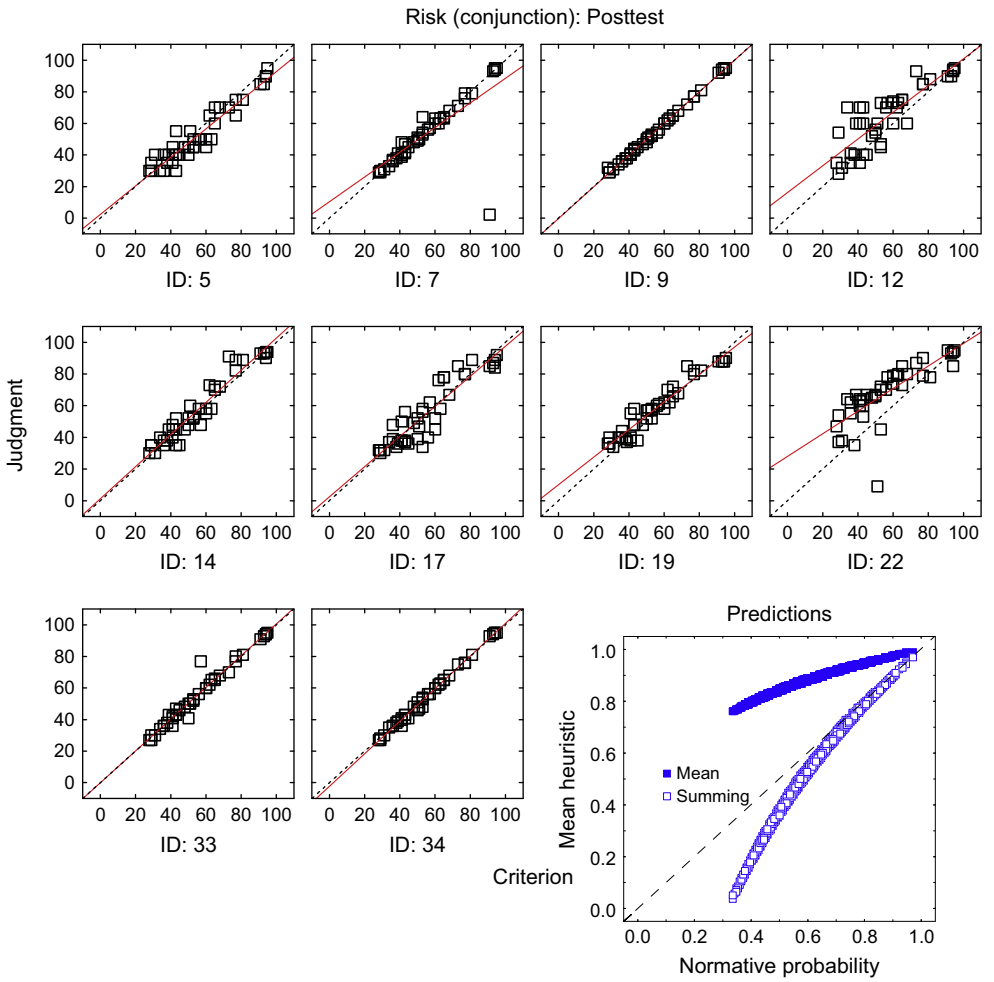

Fig. 6. The judgments by each participant plotted against the normative value in the Posttest with a risk task and a conjunctive frame, together with the predictions by a mean and a summation heuristic (the larger lower panel).

most participants correlated with those predicted by the mean heuristic, but as expected in view of the model fits, this was less so with the risk task.

In sum, the results suggest four conclusions: First, with both task contents and both frames the participants spontaneously addressed the tasks with an additive heuristic, typically the mean heuristic as indicated by the bias in pretest (see Fig. 2).[7] Second, in all conditions the participants improved from a short training by reducing the initial bias. Third, there were significant effects of task content on the judgments. Whereas the conjunctive frame was more difficult than the disjunctive frame with multiple-cue judgment, the performance was similar with both frames with the risk judgments. Only in the posttest with risk judgments in a conjunctive frame, the multiplicative model provided better fit to the data than the additive model. This suggests that the similar performance for risk judgments both in disjunctive and conjunctive frame was explained by improved performance in the conjunctive frame caused by, at least some, participants shifting to use of normative probability combination. With the disjunctive frame, it seems that the participants instead turned to refinements of summation, notably achieving a very similar level of performance. In other words, in the posttests of all of the conditions participants appreciated the interaction between the cues, but the ability to perform multiplicative combination seems restricted to risks in conjunctive frame.

More generally, we interpret the results as suggesting that the participants typically started with the mean heuristic. In the disjunctive frame, additional experience with feedback led to a shift to variations of summation, often across all four cues. In the conjunctive frame, the participants often stayed with the mean heuristic, except that

[7] For space reasons we do not report a detailed residual analysis for the Pretest data. However, in all four pretest conditions the residuals correlate negatively with the normative value across the items ($r = -.895$, $p < .001$; $r = -.919$, $p < .001$; $r = -.355$, $p = .034$; $r = -.490$, $p = .002$) and significantly positively with the residuals predicted by the mean heuristic ($r = .885$, $p < .001$; $r = .909$, $p < .001$; $r = .379$, $p = .022$; $r = .505$, $p = .002$). This is consistent with use of the mean heuristic and, as noted in connection with Fig. 2, the bias also suggests that most of the participants relied on the mean heuristic.

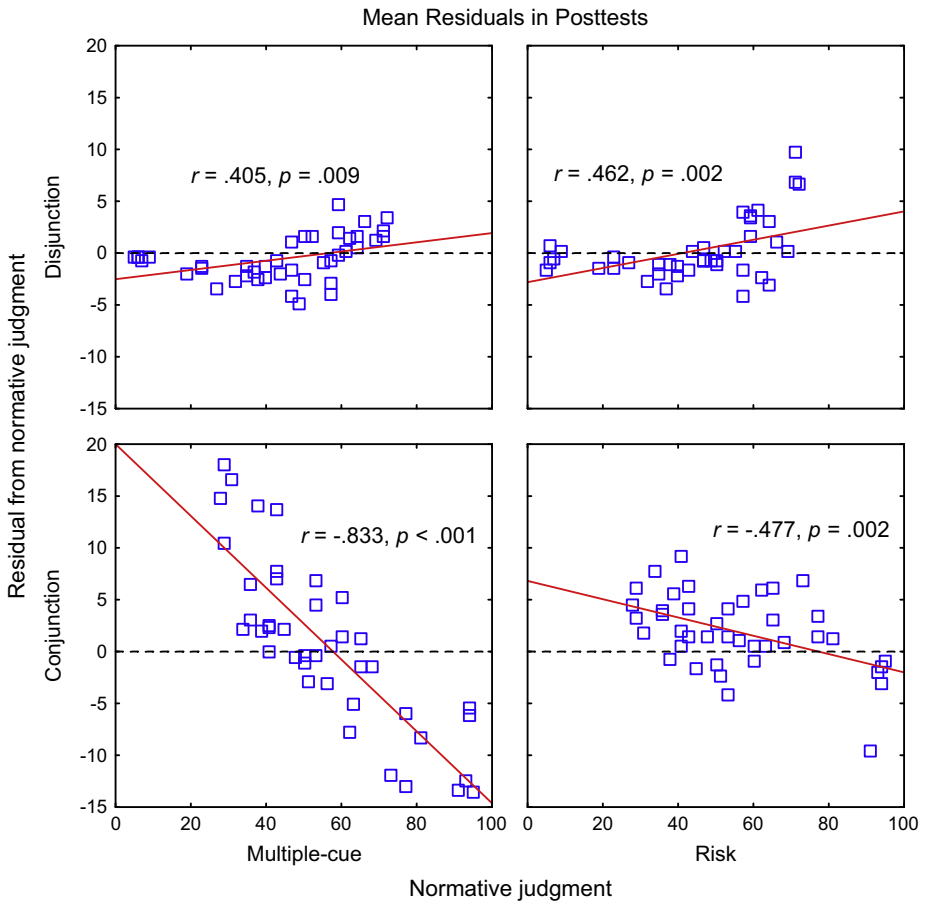

Fig. 7. Mean residuals from the output of the normative (multiplicative) rule, as a function of normative probability, for the posttest of each of the four cells in the design of Experiment 1, together with the correlation coefficients. The two left panels are for the multiple-cue judgments and the two right panels are for the risk judgments. The upper two panels are for the disjunctive frame and the lower two panels for the conjunctive frame.

with risk content some of the participants were able to engage their analytic insights about probability and multiplied.

3. Experiment 2: facilitating detection of the normative rule

In order to make the risk task similar to a standard multiple-cue judgment task, in Experiment 1 there were always 3 or 4 non-zero error sources, thus requiring the combination of several error probabilities/cues. In Experiment 2, we allowed from 1 to 4 non-zero error probabilities. The availability of items with only two non-zero error probabilities should facilitate the detection of the multiplicative rule, especially in the conjunctive frame (e.g., when 1, 1, .9, .9 gives criterion .81). In Experiment 2, we also compared a condition with low criterion probabilities in the training phase to a condition with high criterion probabilities in the training phase, where the latter condition should facilitate detection of the multiplicative rule by highlighting the region where summation and multiplication diverges.

It is also instructive to consider the case of only one non-zero error probability, like, for example, 0, 0, 0, .1 in the disjunctive frame or 1, 1, 1, .9 in the conjunctive frame. If people rely on multiplication or on summation, these items should be especially easy to learn, because the correct answer can be produced without combination, by reporting the probability associated with the non-zero error source (.1, and .9, respectively). However, if people rely on a mean heuristic, these items should not be easier than the other items because they still require the combination of 4 error probabilities. In sum: the aim of Experiment 2 was to facilitate the detection of the normative multiplicative combination rule by introducing items where the detection and execution of multiplication should be simpler and to investigate if the training range for the criterion probabilities affected the ability to detect and use the normative rule. In Experiment 2 we only considered a task content with probability (risk) content.

3.1. Methods

3.1.1. Participants

Participants were 40 undergraduate (16 male and 24 female) students from Uppsala University (M = 22.9 years, Sd = 2.2). They received a movie voucher or course credits as compensation for participating in the study.

3.1.2. Design

Experiment 2 used a $2 \times 2 \times 2$ split-plot design with Frame (conjunction/disjunction) and Training range (high/low) as between-subjects variable and Training (pretest/posttest) as within-subjects variable.

3.1.3. Materials and procedure

Experiment 2 used the same basic procedure as Experiment 1 with a computerized experiment consisting of three parts; pretest, training and posttest. Participants performed the same tasks as in the Risk-condition of Experiment 1 and were asked to assess the probability of an error-free production (conjunctive frame) or the probability of at least one error (disjunctive frame) in the production of a whisky bottle given the probability of no error (conjunctive frame) or the probability of an error (disjunctive frame) in each of four independent production steps (see Fig. 1). Stimulus for the pre- and post-tests was created by orthogonally combining the cue values for each of the four cues (C1: 0, .25, .5; C2: 0, .2, .4; C3: .0, .02; C4: 0, .01) giving a total of 36 items with a criterion range of [0, .71] and [.29, 1.0] in the disjunctive and conjunctive conditions respectively. Four sets of 30 items, two for the high condition and two for the low condition, were created as stimulus for training by randomly drawing error-free probabilities for each of the four cues with the constraint that the total error-free probability should be on the range [.2, .5] in the low condition and [.5, .8] in the high condition. In the low condition cue values were drawn on [0, .1] for all four cues and in the high condition cue values were drawn on [.30, .40] for cue 1 and 2 and [0, .3] for cue 3 and 4. Participants in two conditions were randomly assigned to one of the two training sets for that condition. In pre- and post-test participants judged the 36 items once without feedback and in training participants received feedback on the correct probability after each of the 30 judgments. As in Experiment 1 all four cues were presented simultaneously with cues C1–C4 occupying the production steps A–D respectively and with presentation order within each part of the experiment randomized for each participant.

3.2. Results and discussion

As illustrated in Table 2, which presents the analysis of performance (RMSD) in Experiment 2, the only statistically significant effect was the improved performance from pretest to posttest. The models in Eqs. (5) and (6) were fitted to the individual judgments, separately for the Pretest and the Posttest, in the same way as in Experiment 1. The model fits and the best-fitting parameters are summarized in Table A2 of Appendix. The model fit for both models for each individual participant is presented in Fig. 8. Both the measures for model fit and the parameters were characterized by skewed distributions and heterogeneous variance, requiring the use of on nonparametric statistics.

Fig. 8A and B illustrates that in contrast to in Experiment 1, in the Pretest there were no large differences in the fit of the multiplicative model and the additive model. Fig. 8A and B are more similar to Fig. 4C and D for the Posttest in Experiment 1. There were no statistically significant main effects of Range on the model fit for the multiplicative and the additive models (Mann–Whitney Tests, all $ps > .155$), but the additive model provided better fit than the multiplicative model in the disjunctive condition (Mann–Whitney, $N = 40$, $U = 59$, $Z = 3.801$, $p < .001$). There was no significant difference in the model fit for the multiplicative and the additive model in three out of the four cells (Wilcoxon Test all $ps > .501$), although in the cell with a low training range and conjunctive frame the multiplicative model provided slightly better fit (Wilcoxon Test, $N = 10$, $T = 8$, $Z = 1.987$, $p = .047$). In general, the differences are modest and across all four cells the difference in model fit does not reach statistical significance (Wilcoxon Test, $N = 40$, $T = 343$, $Z = .901$, $p = .368$).

Fig. 8C and D illustrates that in the Posttest there appear to be more consistent differences between the models. There are no significant differences in the model fit for the multiplicative model as a function of Range or Event (Mann–Whitney Tests, both $ps > .113$), no significant effect of Range on the model fit for the additive model (Mann–Whitney Test, $N = 40$, $U = 183$, $Z = .446$, $p = .655$), but a significant effect of Event (Mann–Whitney Test, $N = 40$, $U = 6$, $Z = 5.234$, $p < .001$). The additive model thus provided significantly poorer fit to the data in the conjunctive than in the disjunctive condition. As suggested by Fig. 8C and D, there were no difference in model fit between the multiplicative and the additive models in the disjunctive frame (Wilcoxon Test, $N = 20$, $T = 73$, $Z = 1.194$, $p = .232$), but significantly better fit for the multiplicative model in the conjunctive frame (Wilcoxon Test, $N = 20$, $T = 42$, $Z = 2.352$, $p = .019$). In sum: there was little difference in the model fits, except that in the Posttest with a conjunctive frame, the multiplicative model became superior.

Table 2
Results of a 3-way repeated-measures ANOVA on performance (Root Mean Square Deviation, RMSD) in Experiment 2, where frame and training range are between-subjects variables and training (pretest vs. posttest) is a within-subjects variable with effect size (partial η^2_p). The statistically significant effect is highlighted in bold characters.

	SS	*df*	*MS*	*F*	*p*	η^2_p
Frame	0.0210	1	0.021	1.653	.207	0.044
Training range	0.0125	1	0.012	0.984	.328	0.027
Frame × Training range	0.004	1	0.004	0.293	.592	0.008
Error	0.457	36	0.013			
Training	**0.079**	**1**	**0.079**	**16.008**	**.000**	**0.308**
Training × Frame	0.010	1	0.010	1.925	.174	0.051
Training × Training range	0.004	1	0.004	0.745	.394	0.020
Three-way interaction	0.001	1	0.001	0.200	.657	0.005
Error	0.178	36	0.005			

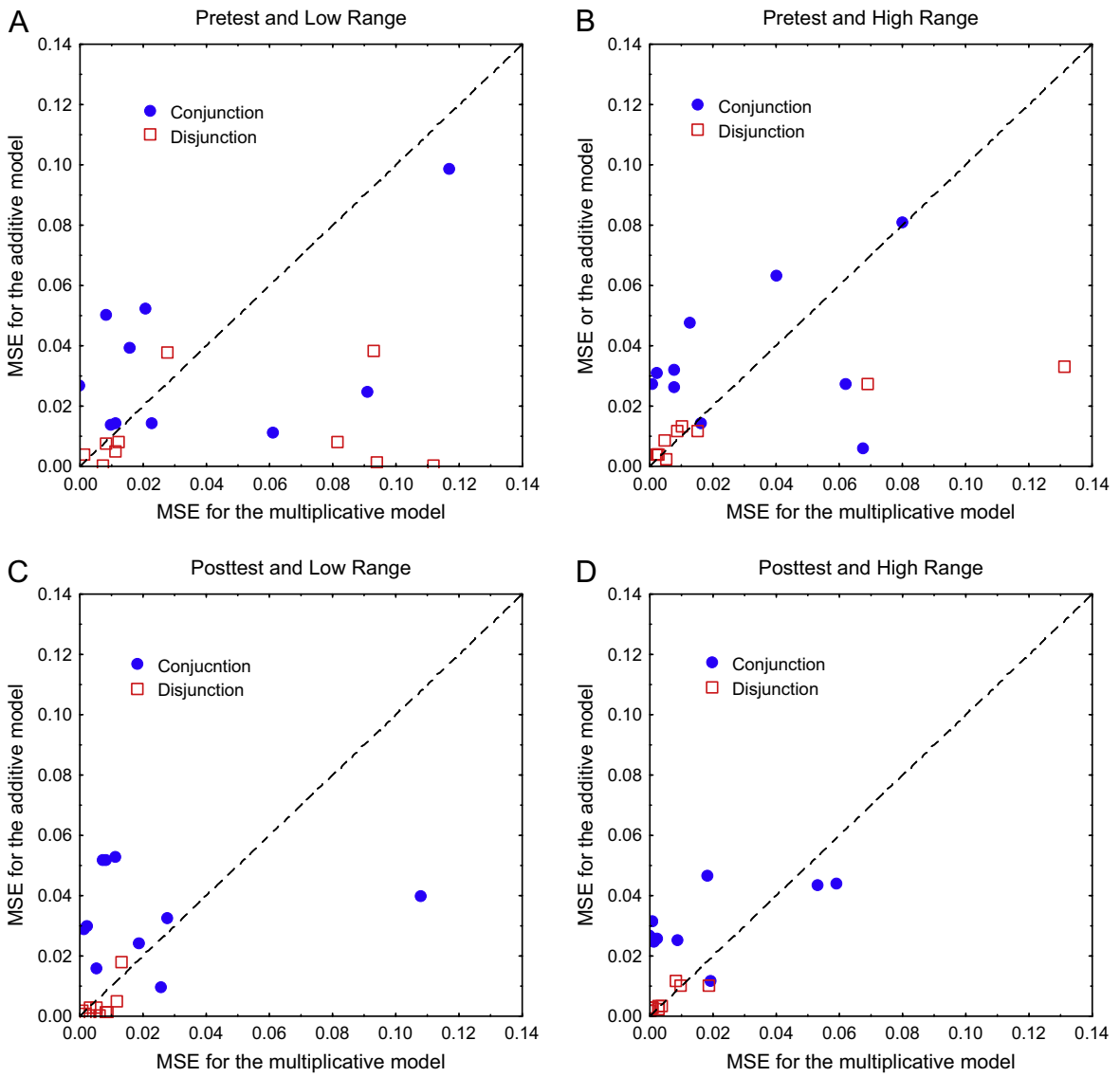

Fig. 8. The model fit in terms of Mean Square Error (*MSE*) between predictions and data for the multiplicative model on the *x*-axis and for the additive model on the *y*-axis, separately for each individual participant in each condition of Experiment 2. Data points above the identity line indicate better fit for the multiplicative model and data points below the identity line better fit for the additive model.

Because there were no significant effects of Range on performance, the two range conditions were collapsed in the residual analysis. We, however, performed separate analyses for each of the two frames (Disjunction vs. Conjunction), considering that Experiment 1 suggested qualitatively different strategies with opposite, and indeed potentially cancelling, biases. Fig. 9 presents the mean residuals from the normative response, separately for the Pretest and the Posttest with both frames, but here plotted as a function not of the criterion but as a function of number of error sources (1,. . .,4) with 95% confidence intervals. Fig. 9 suggest the same qualitative pattern as in Experiment 1. In the pretest with the disjunctive frame (Fig. 9A) there is a significant trend in the mean residuals and all four mean residuals differ significantly from 0, an underestimation consistent with use of a mean heuristic (Fig. 1B). The median correlation between the residuals and the residuals predicted by a mean heuristic across the 36 items at the level of individual participants was .141, thus in the same direction but not significantly different from a population mean of 0 ($t_{19} = 1.383$, $p = .183$, given $H_0 = 0$). In the posttest with the disjunctive frame (Fig. 9B), and despite the improved performance, the mean residuals are still significantly heterogonous with three out of four mean residuals deviating significantly from 0, but now in the direction predicted by use of some modification of summation. The median correlation between the residuals and the residuals predicted by summation across the 36 items at the level of individual participants was .463, significantly different from a population mean of 0 in the direction predicted by summation ($t_{19} = 4.459$, $p < .001$, given $H_0 = 0$). Also as expected if the participants rely on

some variety of summation, the items with one error probability are much easier than those with 3 or 4 non-zero error probabilities, requiring more combination of information.

In the pretest with the conjunctive frame (Fig. 9C), performance is poor with significantly increasing mean residuals, three out of four of which differ significantly from 0. The median correlation between the residuals and the residuals predicted by a mean heuristic at the level of individual participants was .375 ($t_{19} = 3.340$, $p = .003$, given $H_0 = 0$).[8] Performance improves considerably in the posttest (Fig. 9D). In contrast to in the disjunctive posttest, the mean residuals in the conjunctive posttest are not significantly different and three out of four confidence intervals include 0. The median correlation between the residuals and the residuals predicted by a mean heuristic at the level of individual participants was .203 ($t_{19} = 1.255$, $p = .224$, given $H_0 = 0$). The judgments appear about equally biased and variable regardless of the number of non-zero error probabilities, which is suggestive of some strategy that always combines all 4 probabilities. In other words, in contrast to in Fig. 9B, the residuals with one non-zero error source are not more accurate and clearly less variable than the residuals with 4 non-zero error sources. As in Experiment 1, the only condition consistent with a shift toward multiplication is the conjunctive posttest with a risk judgment, while at the same time the participants with a disjunctive frame achieve comparable performance by adapting a summation heuristic.

The results of Experiment 2 suggest three conclusions: First, when the task was simplified by also allowing items with only one or two nonzero error-sources, already in the pretest the participants disclosed appreciation for the interaction of the risks, before they had received any feedback. It is worth contrasting this with the typical result in other multiple-cue judgment tasks, where people often have great difficulty with detecting and implementing cue interactions also after extensive experience or laboratory training (Karelaia & Hogarth, 2008). Second, consistently with Experiment 1, the only condition for which multiplicative combination was supported by the model fits and the residual analysis was the posttest with conjunctive frame. In all conditions, the participants thus revealed a spontaneous qualitative insight that the cues interact in a risk judgment task, but only with a conjunctive frame the data was consistent with performing a multiplication. Third, the performance in this condition was however easily matched by use of a summation heuristic in the disjunctive frame.

4. Experiment 3: with (not against) the cognitive constraints

Experiments 1 and 2 indicate that people have the ability to approximate multiplicative combination of cues by modifications of additive heuristics. Under certain circumstances they are also able to implement multiplicative combination. Nonetheless, it seems that people have great difficulty with implementing truly multiplicative combination of multiple error-sources. In Experiment 3, we therefore tried to aid people's risk combination by framing the task in a format making it possible to arrive at the normative answer by summation, which is better in accord with a predisposition for linear additive combination. In addition, Experiment 3 introduces an analogue format where risks are illustrated graphically. This format is similar to the icon arrays previously shown to improve judgments of risk (Galesic, Garcia-Retamero, & Gigerenzer, 2009) for people low on Numeracy. Thus, the analogue format might serve to facilitate the multiplicative risk combination.

4.1. Method

4.1.1. Participants

Participants were 36 undergraduate (12 male and 24 female) students from Uppsala University ($M = 23.0$ years, $Sd = 3.1$). They received a movie voucher or course credits as compensation for participating in the study.

4.1.2. Design

Experiment 3 used a 3×2 split-plot design with Scale (numeric/analogue/log) as between-subjects variable and Training (pretest/posttest) as within-subjects variable.

4.1.3. Materials and procedure

Experiment 2 used the same basic procedure as Experiment 1 and 2 with a computerized experiment consisting of three parts; pretest, training and posttest. Participants performed the same tasks as in the risk/disjunctive frame condition of Experiment 1. In the numeric condition the risk of an error was presented in the same numerical format as in Experiment 1 and 2. In the log condition the risk was transformed to a logarithmic scale using $p_{\log} = -\log_2(1 - p_{numeric}) \cdot 100$, thus allowing risks to be combined additively. The log scale was presented to the participants as a measure of risk that runs from 0 (no risk) to 999 (practically certain of an error) and that the task was to use the four separate risk estimates to assess the overall risk of an error in the product. Participants were provided with examples of these risk scores and it was emphasized that they were not percentages.

Finally, in the analogue condition the risk was presented in a graphical format where a green horizontal bar was colored red to a proportion corresponding to the risk of each of the four steps and the total risk. Thus a full green bar indicated a 0 probability of an error while a full red bar indicated a probability of an error of 1. In the numeric and log conditions participants gave their answer by entering a numerical value corresponding to the total risk while participants in the analogue condition created a horizontal red bar, by clicking on the appropriate height of a green bar, to indicate their estimate of total risk.

The 36 items used as stimulus for the pretest were created by a uniform random draw of probabilities on the range [0, .5], for the four cues, with the constraint that the total risk should be on the interval [.2, .75]. Risks for each production step were rounded off to multiples of .05. Training used 30 items created by a uniform random

[8] Note that although the residuals predicted by a mean heuristic are a positive function of the number of error sources, as illustrated in Figure 5C, the predicted residuals are a negative function of the criterion, as illustrated in Fig. 1D. This is because with a conjunctive frame, more error sources go with lower probabilities.

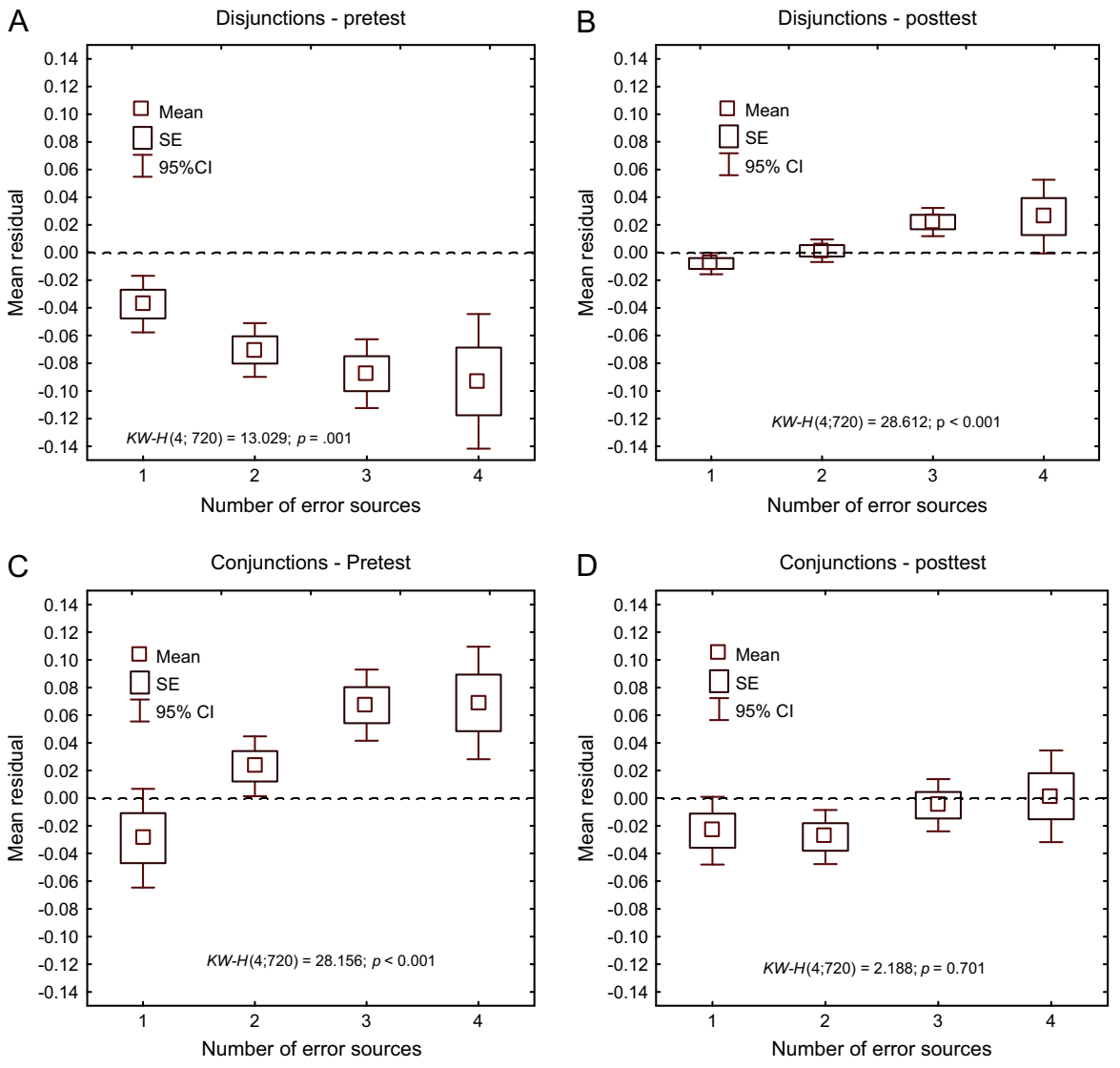

Fig. 9. Mean residuals with 95% confidence intervals from the output of the normative (multiplicative) rule, as a function of the number of non-zero error sources in the judgment item, for the pretest and the posttest and the conjunctive and disjunctive frames in Experiment 2, together with the outcome of a non-parametric Kruskal–Wallis test. The error bars refer to 95% confidence intervals ($N = 20$).

draw of probabilities on the range [0, .4], for the four cues, with the constraint that the total risk should be on the interval [.2, .6]. Finally, the posttest used the 36 items from the pretest and five additional items with a total risk smaller than .1 and thus requiring extrapolation. The cues were presented as in the two previous experiments and the presentation order of items was randomized for each participant within each part of the experiment.

4.2. Results and discussion

Because of the introduction of a log scale in Experiment 3 it proved necessary to analyze the performance (RMSD) with non-parametric statistical tests. Two Mann–Whitney *U*-tests showed no statistically significant difference between the numeric and the analogue conditions, neither in the Pre- or the Posttest ($p > .6$ in both cases). These two conditions collapsed however differ significantly in performance from the condition with the log scale, both in the pretest ($U = 83.000$, $Z = 2.030$, $p = .042$, $N_1 = 24$, $N_2 = 12$) and in the Posttest ($U = 6.000$, $Z = 4.418$, $p < .001$, $N_1 = 23$, $N_2 = 11$). As illustrated in Fig. 10A, performance with the log scale was better both in the Pretest and the Posttest. Performance also improved from the Pretest to the Posttest with both the numerical format (Wilcoxon test: $T = 11$, 3.863, $p < .001$, $N = 23$) and the log format (Wilcoxon test: $T = 6$, 2.191, $p = .028$, $N = 10$).

As illustrated by the mean residuals in Fig. 10B–E, with both formats the participants start off with the bias and the negative correlation between the residuals and the normative criterion predicted by a mean heuristic (Panels 10B to 10E report the correlations between the residuals and the criterion across the items plotted in the figure.) The median correlation between residuals and criterion across the

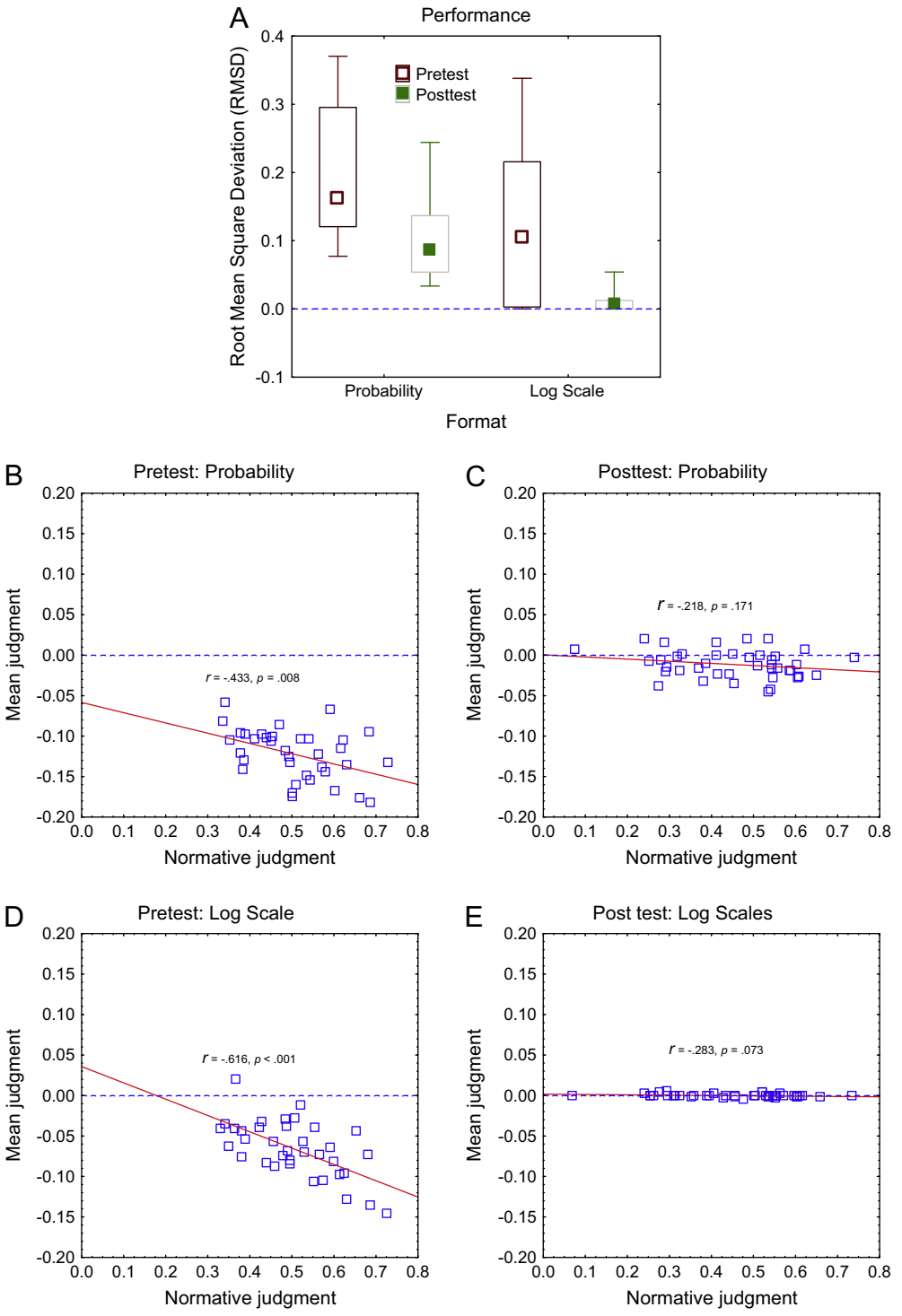

Fig. 10. Experiment 3: Panel A: Median performance in terms of Root Mean Square Deviation (RMSD) from the criterion with Interquartile index (boxes) and min and max range (whiskers) ($N = 24$ for probability and $N = 12$ for the log scale) in the pre- and the posttest in conditions with a probability or a log scale. Panels B–E: Mean residuals from the normative (multiplicative) probability rule, as a function of the rounded normative probability (rounded to 5, 15, 25,. . .,95), for the pre- and the posttest of the two conditions, together with the outcome of a non-parametric Kruskal–Wallis test. The error bars refer to 95% confidence intervals ($N = 24$ for the probability condition and $N = 12$ for the log scale condition).

36 items across individual participants in the pretest with probability scale was −.224, in the direction predicted by the mean heuristic, but not significantly different from a population mean of 0 ($t_{23} = 1.383$, $p = .180$, given $H_0 = 0$). The corresponding median correlation in the pretest with the log scale was −.153 ($t_9 = 3.178$, $p = .011$, given $H_0 = 0$).

In the Posttest with the probability format the participants approximate the normative answers quite well. However, the median correlation between the residuals and the criterion across the 36 items at the level of individual participants in the posttest with probability scale was .272 and significantly different from a population mean of 0 ($t_{23} = 7.422$, $p < .001$, given $H_0 = 0$). With the log scale, in the posttest performance is virtually perfect with close to 0 residuals. The median correlation between residuals and criterion across the 36 items across individual participants in the posttest with log scale was −.031 ($t_9 = .762$, $p = .468$, given $H_0 = 0$). It is clear from a comparison between Fig. 10E and the other figures, in this and in the previous experiments, that the only condition where participants were able to implement the normative combination of risks was with the additive log scale used in Experiment 3.

5. General discussion

Although the traditional experimental paradigms on multiple-cue judgment and probability reasoning tasks share a strong resemblance there are surprisingly few attempts to compare the two tasks directly. In this article, we undertake such a comparison with a special eye to the question of whether the propensity to rely on linear additive combination observed in studies of multiple-cue judgment (e.g., Brehmer, 1994; Juslin et al., 2009; Karelaia & Hogarth, 2008) extends also too formally similar probability combination tasks. Although the different combination strategies often correlate highly in these tasks, and despite the presence of considerable individual differences, by the use of computational modeling and residual analysis we have been able to detect some systematic differences between the tasks.

The results suggest three general conclusions. First, most of the participants seem spontaneously to approach both tasks by using additive heuristics. That is, in the pretest of the experiments most participants combine the available information by means of simple additive heuristics such as summing or averaging, although, as illustrated in Experiment 2, when the task involves risk and items with few error probabilities many of the participants appear to appreciate spontaneously that the risks interact in some way. The overall pattern of bias in the pretests suggests that the most common default strategy is a mean heuristic, which is consistent with previous research multiple-cue judgment (Brehmer, 1994; Juslin et al., 2009; Karelaia & Hogarth, 2008) and probability combination (Bar-Hillel, 1973).

Second, participants nonetheless seem to solve the two types of tasks somewhat differently. With multiple-cue content, the participants adapt to the task by using various modifications of summation, whereas with risk content more participants seem truly to turn to multiplicative combination. The relative performance with disjunctive frame and conjunctive frame in Experiment 1 was therefore different depending on the task content. With multiple-cue judgments, the difficulty of the disjunctive and conjunctive frames accorded with the order of difficulty predicted by a summation heuristic, but not with risk judgments. The only condition in Experiment 1 where both the model fits and the residual analysis were consistent with truly multiplicative integration, was in the Posttest for risks in a conjunctive frame. This suggests that the Task effect was that in the Posttest for risks in conjunctive frame, and in contrast to in the other cells, many participants were able to perform multiplication.

Moreover, with the risk judgments in Experiment 2 the fact that the model fits in Pretest were almost identical and similar to the ones in the Posttest of Experiment 1, suggests that participants had a spontaneous appreciation that the risk sources do interact in some way. It is again worth emphasizing that this does not occur in other multiple-cue judgments tasks, where people often have great difficulty with detecting and implementing cue interactions even after extensive feedback and experience with the tasks (Karelaia & Hogarth, 2008). This suggests that people do have at least some qualitative insight into the need for non-additive combination of multiple risks in a risk combination task, which may be present from start in a task that makes it salient, or that is easily triggered already by a little task experience.

This is consistent with the claim that the mind has little trouble with sequentially adding up or averaging multiple cues, but the sequential and capacity-constrained controlled judgment process has great difficulty with capturing more complex configural cue patterns, except in the simple cases where this process can be amended by direct retrieval of facts from long-term memory (Juslin et al., 2008, 2011). For example, in a multiplicative task where the impact of the individual cues is considered sequentially, the impact of, say, the third cue on the criterion has to be assessed not only in the light of the value of the third cue, but with simultaneous attention to the effect of the first cue and the effect of the second cue.

Our interpretation is that the simple cases where people are able to truly achieve multiplicative combination mainly correspond to items where the product can be produced or well approximated by successive explicit number crunching of two numbers, for example, by retrieving the declarative knowledge that probabilities of independent events should be multiplied together with knowledge of multiplicative facts (the "analytic route" to judgment discussed in Juslin et al. (2011); see Hammond, 1996, for a discussion of analytic judgment). In all other conditions, the residual analysis suggests additive approximations. Many participants appear to approximate normative combination strategy by modifying an additive heuristic, for example, by using proportional summation or curtailed summation.

It is noteworthy how well a risk summation heuristic approximates normative risk combination, especially when the individual risk sources are small (see Fig. 1B). It is worth emphasizing that this good approximation holds already in the case of perfect knowledge of the risk probabilities (i.e., as when the *exact* probabilities are explicitly specified to the participant, as in Fig. 1A). In Juslin et al. (2009) we argued that linear additive approximations to the multiplicative rules of probability theory may be especially useful when the input probabilities to the

combination rule are vague or noisy, as when estimated from small samples, because linear additive combination is often more robust to the effects of random noise than multiplicative combination. In other words, if we in addition take into account that the risk probabilities may often be vague or noisy, it may be even more difficult to detect the "superiority" of the multiplicative combination rule over summation. For many purposes, an agent may be just as well off summing small and noisy risk probabilities.

Finally, only when the risk is represented in a format that makes it possible to combine the error sources by additive operations, as with the log-format in Experiment 3, the participants are able to achieve information combination that is truly normative. This testifies to the fact that, although they do express some qualitative insight into the interaction between the individual risk sources that becomes relevant in a risk combination task, a key constraint on their ability to make perfectly normative judgments refers to the mode of combination.

The position in regard to people's ability to reason with probability advocated here accordingly differs subtly from the standard account in terms of dual process theories (Evans, 2008), where a rapid but fallible intuitive system is supervised and potentially corrected by an analytic system that embodies our normative insights (Kahneman & Frederick, 2002). The results from many experiments suggest that people often do have a qualitative understanding of the probability laws, but that they lack the ability to combine probabilities according to multiplicative rules. People thus in general appreciate that the prior probability (or base-rate) is relevant to the posterior probability of an event (e.g., Koehler, 1996), that, generally speaking, conjunctions tend to be less probable than disjunctions (Nilsson et al., 2009), or that overall risk is a non-additive function of the individual risk components. This qualitative understanding of basic properties of a stochastic environment, which in general is unlikely to take the exact analytic form of the rules in probability theory, may nonetheless be a sufficient basis for applying linear heuristics that allow the mind to approximate the rational learning algorithms captured by Bayesian models of cognition (Oaksford & Chater, 2006). Both the rudimentary qualitative normative insight about stochastic events and the default linear additive combination of information may well operate on an intuitive level and thus be independent of any analytic insights of the form captured by 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, Maria Henriksson, Håkan Nilsson and Anders Winman for valuable comments and discussions of the topics addressed in this article.

Appendix A. Results for model fits and best-fitting parameters

See Table A1 and A2.

Table A1
Median model fit for the multiplicative model (*MSE*(mult.)) and for the additive model (*MSE*(add.)) in terms of Mean Square Error (*MSE*) between predictions and judgments, along with the median best fitting parameters for the multiplicative model (m) and for the additive model (a), for each of the four cells and for the main effects in Experiment 1.

Event condition	Block	Measure	Task condition		Main effect (event)
			Multiple-cue	Risk	
Disjunction	Pretest	*MSE*(mult.)	.027	.051	.039
		m	.889	.843	.889
		MSE(add.)	.007	.011	.010
		a	.810	.575	.626
	Posttest	*MSE*(mult.)	.002	.005	.003
		m	1.048	1.074	1.048
		MSE(add.)	.002	.004	.003
		a	.806	.825	.813
Conjunction	Pretest	*MSE*(mult.)	.074	.058	.070
		m	.437	.706	.547
		MSE(add.)	.023	.013	.015
		a	.187	.189	.187
	Posttest	*MSE*(mult.)	.014	.004	.010
		m	.61	.958	.907
		MSE(add.)	.021	.025	.025
		a	.164	.165	.164
Main effect (task)	Pretest	*MSE*(mult.)	.056	.055	
		m	.550	.706	
		MSE(add.)	.015	.012	
		a	.455	.384	
	Posttest	*MSE*(mult.)	.005	.004	
		m	1.002	1.007	
		MSE(add.)	.007	.016	
		a	.483	.471	

Table A2
Median model fit for the multiplicative model (*MSE*(mult.)) and for the additive model (*MSE*(add.)) in terms of Mean Square Error (*MSE*) between predictions and judgments, along with the median best fitting parameters for the multiplicative model (m) and for the additive model (a), for each of the four cells and for the main effects in Experiment 2.

Event condition	Block	Measure	Training condition		Main effect (event)
			Low range	High range	
Disjunction	Pretest	*MSE*(mult.)	.020	.007	.011
		m	1.020	1.011	1.011
		MSE(add.)	.006	.010	.008
		a	.831	.809	.816
	Posttest	*MSE*(mult.)	.006	.003	.004
		m	1.094	1.027	1.050
		MSE(add.)	.006	.003	.003
		a	.912	.858	.874
Conjunction	Pretest	*MSE*(mult.)	.018	.015	.016
		m	.913	.932	.913
		MSE(add.)	.026	.029	.027
		a	.178	.174	.174
	Posttest	*MSE*(mult.)	.010	.006	.009
		m	.969	.957	.969
		MSE(add.)	.031	.006	.029
		a	.173	.163	.169
Main effect (range)	Pretest	*MSE*(mult.)	.019	.010	
		m	.983	.991	
		MSE(add.)	.014	.020	
		a	.234	.244	
	Posttest	*MSE*(mult.)	.008	.003	
		m	1.065	.997	
		MSE(add.)	.012	.011	
		a	.510	.422	

References

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.

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

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138*, 389–414.

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

Brockner, J., Paruchuri, S., Idson, L. S., & Tory Higgins, E. (2002). Regulatory focus and the probability estimates of conjunctive and disjunctive events. *Organizational Behavior and Human Decision Processes, 87*, 5–24.

Castellan, N. J., & Edgell, S. E. (1973). A hypothesis generation model for judgment in nonmetric multiple-cue probability learning. *Journal of Mathematical Psychology, 10*, 24–222.

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

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58*, 1–73.

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.

Doyle, J. K. (1997). Judging cumulative risk. *Journal of Applied Social Psychology, 27*, 500–524.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment 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.

Evans, J. S. B. T. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. *Developmental Review, 31*, 86–102.

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology, 28*, 210–216.

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

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

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

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin, 138*, 415–422.

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.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General, 108*, 356–388.

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

Jenny, M. A., Rieskamp, J., & Nilsson, H. (2014). Inferring conjunctive probabilities from noisy samples: Evidence for the configural weighted average model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 203–217.

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., Nilsson, H., Winman, A., & Lindskog, M. (2011). Reducing cognitive biases in probabilistic reasoning by the use of logarithm formats. *Cognition, 120*, 248–267.

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.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens 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*, 1–53.

Lagnado, D. A., Newell, B. R., Kahan, S., & Shanks, D. R. (2006). Insight and strategy in multiple cue learning. *Journal of Experimental Psychology: General, 135*, 162–183.

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.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman.

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. A. (2013). Exploring the overestimation of conjunctive probabilities. *Frontiers in Psychology, 4*, 101.

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 & Review, 7*, 375–402.

Oaksford, M., & Chater, N. (2006). *Bayesian rationality*. Oxford: Oxford University Press.

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.

Shaklee, H., & Fischhoff, B. (1990). The psychology of contraceptive surprises: Cumulative risk and contraceptive effectiveness. *Journal of Applied Social Psychology, 20*, 385–403.

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.

Svenson, O. (1984). Cognitive processes in judging cumulative risk over different periods of time. *Organizational Behavior and Human Decision Processes, 33*, 22–41.

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

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

Zacks, R. T., & Hasher, L. (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Betsch (Eds.), *Etc.: Frequency processing and cognition* (pp. 21–36). Oxford, England: Oxford University Press.