Interactivity fosters Bayesian reasoning without instruction

Gaëlle Vallée-Tourangeau, Marlène Abadie, Frédéric Vallée-Tourangeau

Interactivity Fosters Bayesian Reasoning Without Instruction

Word count: 13,058 words

Gaëlle Vallée-Tourangeau

Kingston University

Marlène Abadie

Université de Toulouse (CLLE-LTC)

and

Frédéric Vallée-Tourangeau

Kingston University

Author Note

Gaëlle Vallée-Tourangeau, Department of Psychology, Kingston University; Marlène Abadie, Cognition, Langues, Langage, Ergonomie, Université de Toulouse; Frédéric Vallée-Tourangeau, Department of Psychology, Kingston University.

We thank Nick Shanley for designing the cards used in Experiments 1 and 2. We thank Charissa Bhasi, Niyat Henok, Charles Phillips, and Karis Robinson for their assistance with the recruiting of participants and video data collection and Angie Makri for her assistance with video coding in Experiment 4. We thank Gary Brase, Robin Hogarth, Ulrich Hoffrage, Mike Oaksford, Rob Rehder, Steven Sloman, and two anonymous reviewers for helpful comments and suggestions on an earlier draft of this manuscript.

Financial support from the Kingston University Faculty of Arts and Social Sciences Research Capability Fund is also gratefully acknowledged.

Correspondence concerning this article should be addressed to Gaëlle Vallée-Tourangeau or Frédéric Vallée-Tourangeau, Department of Psychology, Kingston University, Kingston upon Thames, Surrey, UNITED KINGDOM, KT1 2EE. g.vallee-tourangeau@kingston.ac.uk or f.vallee-tourangeau@kingston.ac.uk, tel: +44 (0)208 417 7489, fax: +44 (0)208 417 2388.

Abstract

Successful statistical reasoning emerges from a dynamic system including: a cognitive agent, material artefacts with their actions possibilities, and the thoughts and actions that are realized while reasoning takes place. Five experiments provide evidence that enabling the physical manipulation of the problem information (through the use of playing cards) substantially improves statistical reasoning, without training or instruction, not only with natural frequency statements (Experiment 1) but also with single-event probability statements (Experiment 2). Improved statistical reasoning was not simply a matter of making all sets and subsets explicit in the pack of cards (Experiment 3), it was not merely due to the discrete and countable layout resulting from the cards manipulation, and it was not mediated by participants' level of engagement with the task (Experiment 5). The positive effect of an increased manipulability of the problem information on participants' reasoning performance was generalizable both over problems whose numeric properties did not map perfectly onto the cards and over different types of cards (Experiment 4). A systematic analysis of participants' behaviors revealed that manipulating cards improved performance when reasoners spent more time actively changing the presentation layout "in the world" as opposed to when they spent more time passively pointing at cards, seemingly attempting to solve the problem "in their head". Although they often go unnoticed, the action possibilities of the material artefacts available and the actions that are realized on those artefacts are constitutive of successful statistical reasoning, even in adults who have ostensibly reached cognitive maturity.

Keywords: statistical reasoning, Bayesian inferences, numeracy, systemic cognition, distributed cognition, affordances, flow, task engagement.

Interactivity Fosters Bayesian Reasoning Without Instruction

In contexts where people do not know for sure what the case is or what the future will bring, they still must act, make decisions, and choose between alternatives based on uncertain information and subjective opinions. In court settings, individual jurors must infer the likelihood that the defendant is guilty or innocent based on the accumulation of uncertain pro and con evidence. In medical settings, doctors and nurses must infer the likelihood that their patient has a disease following the observation of the result from a diagnostic test that is susceptible to show a false positive. Ideally, we should be able to reason appropriately with uncertain information. In reality, research has shown that reasoning under uncertainty is often flawed (e.g., Villejoubert & Mandel, 2002) and intervention efforts designed to improve statistical reasoning have met with mitigated success (Kurzenhäuser & Hoffrage, 2002; McCloy, Beaman, Morgan, & Speed, 2007; Villejoubert, 2007).

Over the years, the accumulated evidence suggested that people's use of heuristics was responsible for their poor performance in statistical reasoning tasks (Gilovich, Griffin, & Kahneman, 2002). Heuristic thinking has been attributed to people's general lack of numeracy skills (Chapman & Liu, 2009; Sirota & Juanchich, 2011), lower cognitive abilities, or lack of motivation to engage in effortful thinking (Brase, Fiddick, Harries, 2006; Stanovich & West, 1998). By contrast, in the research presented here, we surmised that individuals' struggle to engage in this type of reasoning, together with researchers' mitigated success in helping participants overcome their difficulties, originates from the type of material commonly used to study statistical reasoning; namely, paper-and-pencil questionnaires. Specifically, we hypothesized that such materials severely constrain what participants can *do* to discover the

correct solution. To test this proposition, we report a series of five experiments showing that performance can be substantially improved when materials afford richer interactions with the statistical information presented in the problems, independently of the information format used and without training. We conclude by discussing how this proposition can help better understand how people's actual thinking capabilities may be realised within and outside the laboratory.

Bayesian Reasoning

Probabilistic reasoning is implicated when one need to *infer* the probability that a hypothesis is true upon receiving new evidence. For example, imagine a head teacher believes or knows *a priori* that there is a 60% probability that a pupil watches too much TV. She then receives a new piece of information about a particular pupil, namely that this pupil needs reading glasses. She should now revise the probability that this particular pupil is watching too much TV. To do so, she needs to consider the chances that a pupil wears glasses if he or she watches too much TV as well as the chances that a pupil wears glasses if he or she watches little TV.

Formally, where H denote the target hypothesis (e.g., H: "the pupil is watching too much TV"), D is the data or evidence received (e.g., D: "the pupil wears glasses"), and $\Pr(H|D)$ is the probability that H is true, given that D has been observed, Bayes's theorem dictates that $\Pr(H|D)$ should be obtained using the following formula:

$$\Pr(H|D) = \frac{\Pr(H) \cdot \Pr(D|\mathrm{H})}{\Pr(H) \cdot \Pr(D|H) + \Pr(\textit{not-}H) \cdot \Pr(D|\textit{not-}H)} = \frac{\Pr(D\&H)}{\Pr(D)} \quad (1)$$

where $\Pr(H)$ and $\Pr(\textit{not-}H)$ represent the prior probabilities that H is true, and that the mutually exclusive, alternative hypothesis, *not-H*, is true, respectively; and where $\Pr(D|H)$ represents the hit rate or conditional probability of observing D if H were true, and $\Pr(D|\textit{not-}H)$, the false alarm rate or conditional probability of observing D if *not-H* were true.

To date, research exploring the nature of the inference processes involved in Bayesian reasoning research typically uses "Textbook problems" (Bar-Hillel, 1983). The following problem provides a typical example (adapted from Zhu & Gigerenzer, 2006):

> The Head Teacher at Teddington School wonders if watching too much TV increases the chances of wearing glasses. He obtained the following information: the probability that a pupil is watching too much TV is 60%. If a pupil is watching too much TV, the probability that he wears glasses is 50%. If a pupil is not watching too much TV, the probability that he wears glasses is 25%. Imagine that a new pupil is wearing glasses. What is the probability that he watches too much TV? ___%

Formally, the problem states that Pr(*H*) = 60%, Pr(*D*|*H*) = 50% and Pr(*D*|*not-H*) = 25% and calls for the value of Pr(*H*|*D*). Applying Bayes's theorem, we find:

$$\Pr(H|D) = \frac{.60 \times .50}{.60 \times .50 + .40 \times .25} = \frac{.30}{.30 + .10} = \frac{.30}{.40} = 75\% \quad (2)$$

In other words, if a pupil is wearing glasses, there is a 75% probability that he is spending most of his free time in front of the television. A substantial research literature has shown that very few individuals can solve such problems when the probabilistic information is presented with single-event probability statements (e.g., "The probability that X is *x*%"): Success rates typically range between 10 to 15% (see Barbey & Sloman, 2007; Koehler, 1996 for reviews).

Human competence for revising prior probabilities in the light of new evidence has long been debated (e.g., Phillips & Edwards, 1966). Kahneman and Tversky's heuristic and biases programme of research was the first to propose descriptive verbal accounts of the heuristic principles people may use to assess uncertainty in general (e.g., the availability heuristic) and posterior probability judgements in particular (e.g., the representativeness heuristic; see Tversky

& Kahneman, 1974). Moving on from these descriptive accounts, researchers have sought to identify means to improve performance. Most notably, research has established it is possible to increase performance considerably by presenting the probabilistic information using "natural frequencies" that provide a summary of frequencies of events, as individuals would have sampled them in their natural environment (Gigerenzer & Hoffrage, 1995). The following version of the glasses problem illustrates this alternative way of presenting the problem data:

> The Head Teacher at Teddington School wonders if watching too much TV increases the chances of wearing glasses. He obtained the following information: 12 out of every 20 pupils watch too much TV. Among these 12 pupils who watch too much TV, 6 wear glasses. Among the 8 remaining pupils who do not watch too much TV, 2 also wear glasses. Imagine you meet a group of pupils who wear glasses. How many of them watch too much TV? ___ out of ___.

Here, the solution is given first by estimating the total number of pupils wearing glasses: there are 8 in total (6 who wear glasses and watch too much TV, and 2 who wear glasses but do not watch too much TV). In other words, 6 out of these 8 pupils watch too much TV. Typically, 40% of participants can solve this type of problems (Barbey & Sloman, 2007).

More recently, studies have also examined the role of individual differences. A robust finding is that high numerate participants—those who have higher abilities for reasoning with basic concepts related to risk and probability (see Reyna, Nelson, Han, & Dieckmann, 2009, for a review)—seem to benefit most from the facilitating effect of natural frequency statements with Bayesian reasoning tasks (Chapman & Liu, 2009; Hill & Brase, 2012; Sirota & Juanchich, 2011). Numeracy levels, however, do not predict the rate of performance with single-event probability statements; performance remains close to zero for both high and low numerate people, indicative

of a floor effect (Chapman & Liu, 2009). So, even if natural frequency statements typically lead to a three to fourfold increase in Bayesian performance (from 10-15% with single-event probability statements to 40-45% with natural frequency statements), there nevertheless remains a majority (typically around 60%) of individuals who do not draw accurate Bayesian inferences.

Various hypotheses have been advanced to account for the facilitating effect of natural frequencies. One view is that the human mind is endowed with cognitive algorithms that are designed to handle frequency information acquired through natural sampling or simply finds it easier to compute frequencies than probabilities (Gigerenzer & Hoffrage, 1995; Kleiter, 1994). Another view is that natural frequencies facilitate performance because, unlike single-event probability statements, natural frequency statements cue a clearer mental representation of the set structure underlying those problems (Barbey & Sloman, 2007; Gigerenzer & Hoffrage, 2007; Girotto & Gonzalez, 2001; Hoffrage, Gigerenzer, Krauss & Martignon, 2002; Macchi, 2000; Sirota, Kostovičová, & Vallée-Tourangeau, 2015). This latter explanation suggests that providing people with a clear presentation of the problem set structure should be sufficient to facilitate the elicitation of correct Bayesian inferences.

External Representations as Cognitive Support?

The ambiguities of a strictly linguistic description of a textbook Bayesian problem must be resolved through interpretation. There is growing evidence implicating the important role of supplementary external representations in addition to the task linguistic description to foster normative inferences in Bayesian reasoning in particular (Brase, 2009; Cosmides & Tooby, 1996; Sedlmeier & Gigerenzer, 2001; Sirota, Kostovičová, & Juanchich, 2014; Sloman, Over, Slovak, & Stibel, 2003; Yamagishi, 2003) and more generally in the study of inductive reasoning

(Vallée-Tourangeau & Payton, 2008; Vallée-Tourangeau, Payton, & Murphy, 2008). Reasoners may represent intermediate aspects of that interpretation in their immediate environment by scribbling down notes, drawing overlapping sets or even some kind of decision tree structure, but such self-generated external representations may or may not include the relevant information to facilitate the required inferences. By contrast, experimenter-generated graphical aids in the form of a non-linguistic external representation of nested sets can offer the explicit segmentation of relevant categories decomposed in terms of countable objects that can be visually inspected (Brase, 2009; Sirota, et al., 2014).

Yet, evidence for the facilitating effect of graphical aids on Bayesian performance is mixed. Using natural frequencies, Cosmides and Tooby (1996) reported that 76% of participants could solve Bayesian problems when the data were also depicted within a grid of 100 squares, with some pre-filled squares. This success rate rose to 92% when participants were first asked to fill in the squares themselves to represent the problem data. These results, however, did not hold when participants were provided with a diagram of Euler circles that depicted relationships between the problem's sets and subsets (Sloman et al., 2003, Experiment 2). The efficacy of graphical aids for improving Bayesian reasoning also appears uncertain with tasks that require combining information that is not segmented in individual cases as with percentage frequency statements such as "*x*% of pupils watch too much TV" or single-event probability statements such as "the probability that a pupil watches too much TV is *x*%".[1] Intensive Bayesian reasoning training programmes lasting between 1 hr 45 min and 3 hrs and using graphical frequency trees as supporting tools can achieve 100% median success rate on tasks using natural frequencies, even 5 weeks after training (Sedlmeier & Gigerenzer, 2001). However, success rate is lower with

less intensive tutorial session (e.g., 50% success rate; see Kurzenhäuser & Hoffrage, 2002) and the benefit of training in one type of problem does not transfer to other types of probability tasks such as cumulative probability tasks (McCloy et al., 2007). Moreover, not all graphical aids can support reasoners: the effectiveness of training with probability trees does not hold up with time (Sedlmeier & Gigerenzer, 2001) while the provision of bar charts may even impede performance when participants are asked to fill the chart themselves (Villejoubert, 2007).

So while graphical aids can be important tools for teaching Bayesian reasoning, the presence of such an aid does not always facilitate performance—or when it does, improvements come at great costs. Yet, so far, we have little understanding of why such diverging effects may occur. Instead, the current view is that most people are simply unlikely to draw Bayesian inferences unless they are endowed with higher cognitive abilities or benefited from top-tier university education (Brase et al. 2006; Stanovich & West, 1998). There is a possible alternative explanation, however. Drawing on the epistemological shift of perspective advocated by proponents of the "distributed cognition" approach (e.g., Hutchins, 2001), we hypothesized that the difficulties most people experience in drawing reasoned Bayesian inferences may not lie in their cognitive limitations or lack of education but, instead, in the tasks and graphical aids researchers have used to evaluate Bayesian reasoning.

Beyond Externalisation: Distributed cognition and Bayesian reasoning

Traditional accounts of human reasoning have situated knowledge and understanding—viz., cognition—within individuals' mind or "inside the skull" as it were. By contrast, an alternative approach to the study of cognition, the so-called "distributed cognition approach" calls for a shift from the mind as the main unit of analysis towards a systemic analysis that encompasses both the

mind, the body, and its surrounding environment (Fioratou & Cowley, 2009; Hutchins, 1995, 2001, 2010; Kirsh, 2009, 2013; Vallée-Tourangeau, in press, 2013; Vallée-Tourangeau, Euden, & Hearn, 2011; Vallée-Tourangeau & Vallée-Tourangeau, 2014; Vallée-Tourangeau & Villejoubert, 2013; Vallée-Tourangeau & Wrightman, 2010; Villejoubert & Vallée-Tourangeau, 2011; Weller, Villejoubert, & Vallée-Tourangeau, 2011; Wilson & Clark, 2009). This entails a reconceptualization of cognition as achieved through the close coupling of internal or mental representations and possible operations together with external or material presentations and possible physical actions people can carry out. Operating within such an extended cognitive system enables people to exceed the capacities of their mental resources since the coupling of physical activity with mental processing augments both the quality and efficiency of thinking. For example, individuals who engage in a difficult mental arithmetic task will be more efficient if they are given the opportunity to manipulate number tokens to support their mental computations as opposed to be constrained to keep their hands still and rely only on their mental powers to compute the sums (Vallée-Tourangeau, 2013). This suggests that levels of performance observed in environments that offer reduced opportunities for coupling thinking with physical actions may be unrepresentative of individuals' true abilities. In other words, the systemic perspective calls for a careful examination, not only of reasoners' mental resources and processing abilities, but also of their immediate material environment and the opportunities (or lack thereof) it offers to support and transform their cognitive efforts.

In classical textbook-problem tasks presenting single-event probabilities statements, as in the vast majority of tasks used to study adult cognition, the immediate environment is often severely constrained: the material presentation of the task consists of a short text printed on a

piece of paper and possibly a blank space where online written protocols are recorded. The tools that can be used to interact with this material presentation are limited to a pen, or perhaps a pencil and an eraser. The material apparatus formed by the printed text, the blank space, and the pen affords the drawing of symbols and self–generated diagrams. These drawings may trigger the use of some learned arithmetic operators and procedures. Being able to see those computations on the blank paper may occasionally increase the effectiveness and accuracy of the computations carried out. However, the overall balance of efforts required to solve these tasks remains heavily skewed towards the mental side, offering limited opportunities to manipulate information in the material world: the problem information cannot be handled, controlled, altered, transformed, or moved through physical action. Descriptive richness or graphical complexity still affords but a restricted range of hands-on manipulations that can meaningfully transform the material presentation of the problem information. Participants must rely on their mental representation of the structure of the task while regulating their thinking process in order to arrive at an answer. This in turn is likely to result in a hefty working memory load, which constrains the cognitive operations that can be applied to the task. Clarification of the nested-set structure of the task (e.g., presenting the problem data in the form of natural frequencies or presenting visual aids) may therefore help because it reduces the mental efforts required to operate on internal representations, thereby allowing those with sufficient cognitive resources to solve the task. For most people, however, such interventions seem insufficient to alleviate the cognitive load imposed by these tasks.

Thus, the performance ceiling observed in Bayesian tasks may have more to do with the impoverished external resources available to support thinking than to participants' cognitive

deficiencies or conceptual limitations. Crucially, this analysis suggests that participants should be able to overcome this performance ceiling if their immediate environment enabled the judicious coordination of mental activity with modifiable material resources, resulting in the development of a more productive problem representation from which they could draw Bayesian inferences. We do not argue that performance is unrelated to individual characteristics in a distributed environment. Rather, we contend that the importance of individual differences (i.e., the cognitive processing individuals are capable of implementing, given their cognitive skills and knowledge) will depend on the level of mental efforts required by the task provided *as well as* the extent to which this task affords the distribution and coordination of efforts between individuals' mind and their immediate environment (i.e., the action possibilities or "affordances" of task materials at the disposal of reasoners when they attempt to solve the task). We now report five experiments designed to explore this proposal.

Experiment 1

In this experiment we aimed to engineer an interactive thinking context in which the information relevant for a Bayesian inference could be observed, counted, manipulated, ordered and reordered to form different categories. We purposely did not instruct participants on how to combine information or how to apply Bayes's theorem. Instead, we provided them with a pack of custom-made playing cards depicting the individual elements of the sets described in the problem (Figure 1 illustrates the cards accompanying the glasses problem presented earlier). We surmised that the physical activity that these cards afforded and the associated dynamic perceptual feedback would recruit a broader range of perceptual and cognitive processes in solving the problem compared to those implicated in the interpretation of standard linguistic and

diagrammatic problem descriptions. Thus, by allowing participants to recruit and coordinate internal and external resources, we anticipated that they would develop a richer and more complex mental representation of the data that might convey more transparently key steps in deriving the correct Bayesian answer. In line with previous research findings (Chapman & Liu, 2009; Hill & Brase, 2012; Sirota & Juanchich, 2011), we also anticipated that individuals with higher numeracy skills would outperform those with lower numeracy skills.

Method

Participants. A total of 90 individuals (64 women and 26 men; mean age = 22 years, *SD* = 5.66) volunteered to take part in the experiment. The data were collected individually at a public library. The vast majority (94.4%) were Social Science and Humanities undergraduate students. The experiment was conducted in French.

Design and procedure. Participants were randomly allocated to one of two conditions: a high-interactivity condition (paper-and-pencil with cards) or a low-interactivity condition (paper-and-pencil only). We used three scenarios adapted from Zhu and Gigerenzer (2006): the glasses problem presented above, the student problem and the cat problem. The probability data were presented using natural frequencies, with three sets of data (see Table 1). Scenarios and sets of frequency data were rotated to create 9 versions of the questionnaire, which were randomly allocated to participants.

All participants received a 2-page questionnaire presenting the three problems to complete on the first page. A blank space of approximately 5 cm had been left between each problem to allow participants to record their thoughts. The second page presented a French translation of the numeracy scale developed by Lipkus, Samsa, and Rimer (2001; See Appendix).

Participants in the high-interactivity condition were also provided with three sets of 4" x 2.5" custom-made cards representing the elements in each three problems presented (see Fig. 1 for an illustration). Thus we used a 2 (level of interactivity) × 3 (problem) design, with repeated measures on the last factor.

Participants were approached by the experimenter who invited them to take part in a study on their ability to reason with numbers. Upon consenting to participate, people in the low-interactivity condition received the questionnaire and were asked to use the space below each problem to explain how they arrived at their answer, either by writing down all the steps in their reasoning or by noting which numbers they used for deriving their answer.

Participants in the high-interactivity condition also received three packs of 20 cards, one for each problem. Each pack contained a number of *H* & *D*, *H* & *not-D*, *not-H* & *D* and *not-H* & *not-D* cards which matched the quantities stated in the corresponding problem. These cards were introduced by telling participants that the problems they were about to solve were quite difficult and that using cards had been shown to help solving them. The experimenter explained that she was to observe how they would use the cards for solving the problems. The experimenter then presented the pack of cards corresponding to the first problem on the questionnaire and explained that the cards represented the different possibilities mentioned in the problem. The following script illustrates how participants who had to complete the glass problem first were introduced to the cards:

> "For example, the first problem (*experimenter picks the relevant pack of cards*) relates to leisure activities such as watching television (*experimenter presents a card showing a television*) or riding a bicycle (*experimenter presents a*

card showing a bicycle). Some of these cards show a child with glasses on the back while other show a child without glasses. For example, if there is a television on one side, there might be a child with glasses on the other side (*experimenter turns a television card and reveals a child with glasses*) or a child without glasses (*experimenter turns another television card over and reveals a child without glasses*)."

Participants were instructed to arrange the cards so that they could better understand the information presented in the problems. For each problem, all cards were shuffled and presented in a deck with information about one of the base-rate categories (e.g., a television or a bicycle in the glasses problem) facing up. After having solved all three problems and completed the numeracy scale, they were thanked and debriefed.

Results and Discussion

Answers were classified as Bayesian using Gigerenzer and Hoffrage's (1995) strict outcome criterion. Specifically, an answer was categorized as Bayesian if the numerical response matched the Bayesian solution perfectly (rounded up or down to the next full percentage point), for a given set of numbers (see Table 1). Moreover, written protocols were used to classify answers—answers from participants who provided a fraction without performing a division or who produced a correct Bayesian ratio but made a calculation error for the final division were nevertheless classified as Bayesian.

The primary objective of this study was to assess whether the level of interactivity afforded by the task would affect Bayesian performance. We also wanted to examine whether numeracy would moderate the effect of interactivity and whether performance would increase

with practice. Numeracy scores on the Lipkus et al.'s (2001) were skewed towards the maximum scores (median = 9, range = 3–11, Cronbach's α = .610). This is a well-documented issue with this scale, which has usually been addressed in past research by using median splits (e.g., Chapman & Liu, 2009; Hill & Brase, 2012; Peters et al., 2006). Median splits in multiple predictor models, however, can also create spurious effects by increasing the probability of Type I errors, especially for interaction tests (Maxwell & Delaney, 1993). Thus, to avoid the issues associated with using median splits for continuous predictors, we examined the effect of numeracy, practice, and interactivity using a model comparison approach (Judd, McClelland, & Ryan, 2009).

To test for between-subject effects, we regressed the average performance over the three problems on three predictors: interactivity (contrast-coded -1 for paper-and-pencil and 1 for paper-and-pencil with cards), the mean deviation form of the numeracy score, and the product of these two variables. We regressed the difference in performance between problem 1 and 3 on these predictors to test for within-subject effects. The descriptive statistics and intercorrelations for the model variables are presented in Table 2. The results of the regression analyses are presented in Table 3.

People were able to draw Bayesian inference, as the overall mean proportion of Bayesian answers was significantly different from zero. The increase in interactivity offered by the cards was associated with a significant increase in performance: $M_{\text{Low\_interactivity}}$ = .52 , SD = .41, $M_{\text{High\_interactivity}}$ = .73, SD = .36, p = .008. Higher numeracy scores also resulted in better performance, p = .009, as did practice; $M_{\text{First\_problem}}$ = .56, SD = .50, $M_{\text{Third\_problem}}$ = .69, SD = .47, p = .013. There was no evidence that numeracy moderated the effect of interactivity on

performance, $p = .30$, the effect of practice, $p = .748$, or both, $p = .549$. Interestingly, however, the interaction between interactivity and practice was statistically significant, $p = .036$. Figure 2 illustrates this finding. Within-subject contrast tests for the role of interactivity on practice revealed a significant linear trend between practice and performance in the high interactivity group, $F(1, 44) = 14.24$, $p < .001$, but not in the low interactivity group, $F < 1$. So, independently of numeracy skill levels, higher interactivity (pen-and-pencils with cards) resulted in improved performance through practice whereas performance stagnated in the low interactivity context (pen-and-pencil only). These results thus demonstrate that providing participants with the opportunity to interact with the problem data through sampling cards greatly enhances their performance, independently of their numeracy skills.

Experiment 2

In Experiment 1, we tested the effect of interactivity using Bayesian reasoning tasks that presented statistical information in the form of natural frequency statements. This information format is known to facilitate Bayesian performance (e.g., Gigerenzer & Hoffrage, 1995). By contrast, as we reviewed earlier, problems that use single-event probabilities statements are notoriously harder to solve: performance rate usually plummet to 10 to 15%, and intensive training sessions are required to help individuals draw appropriate inferences from tasks using this information format. Experiment 2 aimed to investigate whether performance would still benefit from the increase in interactivity afforded by the availability of playing cards when using single-event probability statements to present the statistical data. Specifically, it was designed to test which of the following three alternative predictions held true. First, a strong prediction derived from the systemic cognition approach would be that allowing participants to coordinate

mental and material resources should be sufficient to enable them to draw Bayesian inferences, even without instructions, and independently from the statistical information format. If this were the case, the majority of participants should be able to solve problems using single-event probability statements in an interactive thinking context. Second, a weak prediction would be that the manipulability of the task information is a necessary but not sufficient condition for enabling Bayesian reasoning: there also needs to be a one-to-one mapping between the parameters that define the task at the abstract level (e.g., natural frequencies) and the parameters that define the task at the concrete level (e.g., countable cards). If this were the case, merely providing participants with countable cards in addition to the linguistic description of a Bayesian task using single-event probability statements should not be sufficient to elicit correct judgments from the majority of participants. Finally, an intermediate prediction would be that participants who have sufficient mental resources (e.g., in the form of higher numeracy skills) would be able to use the manipulable materials as a scaffold whereas those with lower cognitive resources would not.

Method

Participants. A total of 90 individuals (58 women and 32 men; mean age = 23 years, *SD* = 4.28) volunteered to take part in the experiment. The data were collected individually at a public library. The sample included 79% Art and Social Science students and 10% of Mathematics or Science students. Participants were either post-graduates (47%), undergraduates (48%), or had completed high school or did not specify (5%). The experiment was conducted in French.

Design and procedure. As in Experiment 1, participants were randomly allocated to one of two conditions: a high-interactivity condition or a low-interactivity (control) condition and completed 3 problems, thus a 2 (level of interactivity) × 3 (problem) design was used, with repeated measures on the last factor. The materials used were identical to those used in Experiment 1, except that the numerical data in all problems were presented in the form of single-event probability statements (e.g., "The probability that a pupil watches too much TV is 60%"; see Table 1 for the full set of numerical data used). Participants' strategies for computing the final answer as well as their score on the translated Lipkus et al.'s (2001) 11-item numeracy scale were also recorded.

Results and Discussion

Answers were again classified as Bayesian using a strict outcome criterion, as in Experiment 1. The primary objective of the present experiment was to examine whether the increase interactivity level afforded by the cards would also improve Bayesian performance when the problem information was based on single-event probabilities statements. A secondary objective was to examine whether practice and numeracy moderated the effect of interactivity on performance. We subjected the data to the same model comparison analysis used in Experiment 1. The descriptive statistics and intercorrelations for the model variables are presented in Table 4. The results of the regression analyses are presented in Table 5.

As in Experiment 1, there was evidence that people have the ability to draw Bayesian inferences, as the overall mean performance was significantly different from zero. Once more, higher interactivity levels led to a significant and substantial increase in performance rate; $M_{\text{Low\_interactivity}} = .09$, $SD = .25$, $M_{\text{High\_interactivity}} = .57$, $SD = .42$, $p < .001$. The only other

significant predictor of performance was the level of numeracy, $p = .02$, although numeracy did not moderate the impact of interactivity. Unlike what we observed in Experiment 1, practice did not significantly improve performance. Those results thus show that increasing the level of interactivity afforded by the Bayesian tasks was sufficient to enable the majority of participants to draw accurate statistical inferences even when the statistical information is presented in the form of single-event probability statements: 58% of participants successfully solved the last problem in the high interactivity condition, compared to 9% in the low interactivity condition. This finding thus supports the strong prediction derived from the systemic cognition approach. The fact that practice no longer led to a linear improvement of performance with increased interactivity suggests that participants' cognitive resources were stretched by the coupling cost involved in mapping the problem information (presented in the form of single-event probabilities) and its material implementation (represented in the form of playing cards). This is evidenced by the fact that the success rates observed in this experiment were much lower than those achieved in Experiment 1 with natural frequency statements where, for example, 84% of participants correctly solved the last problem in the high interactivity condition. This suggests that natural frequencies also have a facilitating effect, above interactivity.

There are possible limitations to these results, however. First, the experimental conditions in Experiments 1 and 2 offered higher levels of interactivity and afforded more action possibilities than the controlled condition, but this was not the only difference. The use of playing cards in the high interactive conditions also led to the specification of information that remained implicit in the classic paper-and-pencil versions of Bayesian reasoning tasks. For example, in the glasses problem, the use of cards resulted in the explicit description of pupils

who did not watch too much TV as bicycles riders (in the form of a bicycle image on the cards in the high-interactivity condition). So whereas the classic paper-and-pencil version of the problem mentions that "12 out of every 20 pupils watch too much TV," the provision of cards also specified the contrast to the base category as it represented this information using a sample of 12 cards showing a television as well as 8 cards showing a bicycle. To test for this potential confound, Experiment 3 examined whether explicitly unveiling the alternative base rate category in a paper-and-pencil task would be sufficient to replicate the level of performance observed with cards.

A second limitation concerns the reasoning processes that may underpin performance in highly interactive conditions. One could argue that successful participants did not 'reason' through the task but 'simply' counted the number of cards presenting the target hypothesis *H* among the cards presenting the data *D*. Such a strategy, however, is only seemingly simple. In the glasses problem mentioned above, this strategy would amount to (1) *sorting* the deck of cards into a pile of cards showing *D,* a child wearing glasses, and a pack of cards showing not-*D,* a child without glasses; (2) *counting* the number of cards showing a child with glasses, *n*(*D*); (3) within the set of cards showing *D,* a child without glasses, *sorting* the cards into a pack of cards showing *H*, a television on the other side and a pack of cards showing *not-H*, a bicycle on the other side; (4) *counting* the number of cards showing a television among the cards showing a child with glasses on the other side, *n*(*D*&*H*); (5) *report n*(*D*&*H*) out of *n*(*D*). In Experiment 2, the experimenter made notes of the different manipulations that spontaneously arose in the high-interactivity condition for a subset of participants ($n = 27$). Table 6 summarises these observations. A little under half the participants who correctly solved the problem worked

through the problem by sorting cards in *D* and not-*D* piles. The remainder started from the base-rate information, and began by sorting cards in *H* and *not-H* piles. It is difficult to argue that one strategy is superior to the other. Sorting cards based on the presence or absence of *D* is slightly more efficient—participants can report the solution in 4 steps, compared to 5 steps when one starts by sorting *H* and *not-H*. Nevertheless, these data strongly suggest that participants did engage in thinking and reasoning about the task, even if they give only a cursory snapshot of the actions that participants actually performed while progressing towards the task goal.

A third limitation of these experiments concerns the possible generalizability of the effect of interactivity. The sets of cards provided to support reasoning were such that they always provided an accurate presentation of the probabilities in the problems. For example, when the problem stated that "Among these 12 pupils who watch too much TV, 6 wear glasses", the sample of cards given to participants contained exactly 12 cards showing televisions on the front side and, among those 12 cards, 6 revealed a child with glasses on the backside and the remaining 6 revealed a child without glasses. It is therefore unclear whether the provision of cards would continue to boost participants' performance (and to what extent) in less constraining circumstances, such as in the absence of a perfect match between the statistical data in the problem statement and the sample of cards provided. In like vein, the extent to which those results depend on intrinsic features of the cards we have used (i.e., providing information printed on opposite sides) also bears questioning. Experiment 4 provides an empirical test of these issues.

Experiment 3

Method

Participants. A total of 70 Psychology Students (58 women, 11 men, 1 unspecified, mean age = 24 years, *SD* = 7.85) took part in the experiment in exchange for course credits. The data were collected in a classroom setting. The experiment was conducted in English.

Design and procedure. Half the participants received a problem with natural frequency statements, whereas the other half received a problem using single-event probability statements. Within each of these conditions, half of the participants received a standard version of the problem specifying Pr(*H*), Pr(*D*|*H*) and Pr(*D*|not-*H*); for the other half, the problem fleshed out all probabilities and thus explicitly mentioned Pr(not-*H*), Pr(not-*D*|*H*), and Pr(not-*D*|not-*H*) as well. We used six sets of statistical data and three sets of scenarios. Participants were randomly allocated one version of the resulting sample of 72 problems.

Results and Discussion

Answers were classified as Bayesian using Gigerenzer and Hoffrage's (1995) strict outcome criterion. None of the participants found the correct answer with the problem using single-event probability statements, whether it presented the standard information (n =17) or the fully fleshed out data (n = 18). Using natural frequencies helped improve performance, 41% of participants responded with the Bayesian answer in the standard natural frequency version of the problem. This success rate dropped to 28% using the fully fleshed out natural frequency version although this decrease in performance was not significant, $\chi^2(1, N = 35) = .70$, $p = .41$. Thus, there was no evidence that fleshing out the statistical information in a paper-and-pencil problem would improve performance. If anything, the evidence suggests it might impair it, possibly

because fleshing out implicit information gives rise to a more complex representation but, by the same token, increases the cognitive costs involved in maintaining it in working memory. These results contrast with the sizeable improvement in performance observed with the playing cards and suggest that the cognitive support provided by the cards goes above and beyond the provision of an explicit representation of the alternative base rate category.

Experiment 4

Experiment 4 was designed to assess the generalizability of the facilitating effect of interactivity over problems whose numeric properties do not map perfectly onto the cards, over the form of the material presentation provided, as well as to further investigate how the manipulation of playing cards may give rise to successful statistical reasoning. More specifically, this experiment had three objectives. First, it aimed to examine whether reasoning performance would still benefit from the provision of playing cards when the sample of cards provided was not isomorphic to the sample described in the problem text. In this situation, participants would first need to select the appropriate number of cards from each subset *H&D, not-H&D, H&not-D, not-H&not-D* to recreate the set described in the problem. Pilot testing revealed that upon realizing the set of cards provided was non-isomorphic to the problem data, participants simply stopped interacting with the cards and attempted instead to solve the problem "in their head". At debriefing, they explained they had inferred the experimenter had "tricked them" into using the cards and concluded they were instead expected not to use them, despite being explicitly instructed to do so to solve the task. To circumvent this issue and nevertheless examine whether handling cards may support performance even using non-isomorphic samples, we used a within-subject design where participants first interacted with sets of cards that were isomorphic to the

sets described in the problem text, followed by trials where the sets of cards and the sets described in the text were no longer isomorphic. This had the advantage of allowing participants to learn independently (i.e., without instructions) how to couple the cards to the written information and next to examine whether such coupling would continue to benefit their thinking even when the card sets and the problem data were no longer aligned.

A second objective was to examine whether statistical reasoning performance would still be improved by the provision of manipulable playing cards to represent statistical information when information about the hypothesis and the data was printed side by side on the front side of cards. On the one hand, one might expect that providing information side by side would not alter performance if improvement were underpinned by the increased manipulability of the material afforded by the playing cards. On the other hand, making information about both the hypothesis (*H* or *not-H*) and the data (*D* or *not-D*) available on one side of the card might increase processing costs since it requires reasoners to consider two pieces of information at once instead of one. This increase in processing cost could dampen the rise in performance originally observed with the two-sided cards we had used in previous experiments. Given this possibility, and to avoid floor effects, we chose to present the problem data using natural frequency statements since all participants benefitted from the use of cards in Experiment 1, whether they had high or low levels of numeracy.

Finally, the third and final objective of this experiment was to provide a better understanding of how interactivity worked while participants used the cards to support their thinking by capturing and analysing their actions as they unfolded in time, using systematic observation and sequential analysis (Bakeman & Quera, 2011).

Method

Participants. A total of 20 Psychology Students (18 women, 2 men, mean age = 25 years, *SD* = 7.66) took part in the experiment in exchange for course credits. The data were collected individually in the Kingston Psychology Observation Laboratory. The experiment was conducted in English.

Design and procedure. Participants were invited to solve a series of six Bayesian reasoning problems presenting the statistical information using natural frequency statements. We used six scenarios adapted from Zhu and Gigerenzer (2006): the three scenarios used in Experiments 1 and 2 as well as the cookie problem, the teeth problem, and the overweight problem. We used six new sets of data (see Table 7). Scenarios and sets of frequency data were rotated to create 24 versions of the questionnaire, which were randomly allocated to participants.

Upon consenting to participate and to be filmed, participants were sat at a table. The experiment unfolded in two parts. In the first part, participants were asked to solve three Bayesian problems with the help of an associated pack of cards. The packs were prepared so that the number of cards provided matched the frequency counts in the problem statements (isomorphic samples). The experimenter introduced the cards for the first task with the instructional script used in Experiment 1. In addition, participants were told that they would need to make use of all the cards given to them to better understand the information presented in the problems. Each problem statement was printed on an A4 sheet page using a 26-point font size and was accompanied by a pack of 4” x 2.5” custom-made cards representing the elements in each problem side-by-side (see Fig. 3 for an illustration). In the second part, participants were asked to solve an additional three sets of problems but were also informed that, this time, they

would first need to decide how many cards they should use to better understand the information presented in the problems. They were provided with packs of cards containing ten exemplars of each card type (*D&H, D&not-H, not-D&H*, and *not-D&not-H*). While participants were working on the problem, the experimenter returned to a control room and filmed their hand movements from an overhead camera. Upon each task completion, participants called the experimenter and announced their answer. They were then provided with the next task statement and associated pack of cards without further instruction. After the third task was completed (end of Part 1), participants filled in Lipkus et al.'s (2001) 11-item numeracy scale. When they had announced their solution to the sixth problem, participants were thanked and debriefed. Thus, a 2 (set correspondence) × 3 (problem position) within-subject design was used.

Results and Discussion

Bayesian performance. Answers were classified as Bayesian using a strict outcome criterion as in previous experiments. The primary objective of the present experiment was to test whether the increase in performance observed when participants could use playing cards to support their thinking was due to the fact that the number of cards matched the statistical information in the written description of the Bayesian tasks. We also examined whether practice and numeracy moderated the effect of interactivity, using a different set of playing cards showing data and hypothesis information printed side by side. We subjected the data to the same model comparison analysis used for the data from Experiment 1. The descriptive statistics and correlations for the model variables are presented in Table 8. The results of the regression analyses are presented in Table 9.

There was no evidence that the type of sample (isomorphic vs. non-isomorphic) made a difference to participants' Bayesian performance; $M_{\text{isomorphic}} = .58$, $SD = .36$, $M_{\text{non-isomorphic}} = .58$, $SD = .47$, $p = 1.00$; likewise, numeracy did not affect performance, $p = .272$. Practice, however, did have a significant effect on performance, $M_{\text{first}} = 0.45$, $SD = 0.43$, $M_{\text{last}} = 0.68$, $SD = 0.41$, $p = .016$, thus replicating the practice effect associated with the use of cards we had observed in Experiment 1. None of the interaction terms reached statistical significance.

Behavioral analysis. To provide a better understanding of the processes by which an increased level of interactivity may support participants while they worked through the problems, we recorded and analyzed their hand movements from one camera attached to the ceiling and one camera attached on the wall of our observation lab (see Fig. 4 for an illustration of the videographic evidence). To shed light on the qualitative differences between high and low numerates, we selected the recordings of high numerates who correctly solved the first three tasks ($n = 21$) and the recordings of low numerates who did not succeed at solving the first three tasks ($n = 16$). We restricted our analysis to the isomorphic problems to ensure that we only coded behaviors related to representation and information processing as opposed to efforts to transform the non-isomorphic sample of cards in an isomorphic one. Following Bakeman and Quera (2011), we began to develop a coding scheme through an iterative process. We first watched a few video recordings repeatedly to identify generic mutually exclusive and exhaustive codes for behaviors (e.g., "does not touch cards", "picks up the pack of cards", "examines cards", "moves cards").

We then pilot-tested our coding scheme on new video recordings to identify codes that could be reliably applied to all videos. Coding was also theoretically grounded in concepts from

the literature on distributed cognition and insight problem solving. This iterative process led us to define four types of activities participants engaged in: Projection, Marking, Presentation change, and Epistemic activity. Each type of activity was defined in terms of specific actions we observed. Table 4 provides details of the final coding scheme. The "projection" activity makes reference to the process of projecting mental representations onto the visible environment hypothesized by Kirsh (2013). We coded this activity whenever participants made no action on the cards. The "marking" activity refers to a type of behavior known to support cognition by directing attention and helping perception (Carlson, Avraamides, Cary, & Strasberg, 2007; Kirsh, 1995). We coded this activity whenever participants interacted with the cards without gathering information or making significant change to the perceptual layout. Marking actions included nudging the cards slightly, marking cards with a hand, or one or more fingers, holding onto the cards. The "presentation change" activity was informed by the concept of representation restructuring in the insight problem-solving literature, which describes reasoners' attempt to restructure the representation of a problem when searching for a fruitful solution (Fleck & Weisberg, 2013). Presentation change actions included picking up, putting down or laying cards out and transforming the layout by rearranging cards. Finally, the "epistemic activity" included actions aiming to support mental computations or uncover information that is hidden (Kirsh & Maglio, 1994), such as sampling the cards and counting. To ensure that all actions coded were mutually exclusive, we also coded quantitative cues. For instance, if a participant moved a card less than 2cm away from its original position on the table, this was coded as a "Nudges cards" action (Action 1.1 in the Coding Scheme). The behaviours observed in the video recordings were coded using the Noldus Observer XT 11.0 software[2] to record onset and offset times of the

behavioral events listed in Table 10, continuously sampling behavior from the beginning of the task until participants announced their answer. Table 11 illustrates the data obtained by coding the video footage for one participant.

To evaluate inter-coder reliability, we trained a research assistant, blind to the outcome of the trial (successful or unsuccessful) to code the videos using the coding scheme developed by the first author (Table 11). The initial average Cohen's κ, taking both the type of action and the sequence of events coded into account with a tolerance window of 1s, was .78, with an 82% average percentage of initial agreement. All disagreements were resolved through consensus. The final Cohen's κ taking both the timing of coding and the sequence of events into account was .88, with a 91.35% average percentage agreement.

We then analyzed two behavioral measures: the total amount of time participants engaged in each category of behavior (in seconds) and the proportion of time they did so out of the total time they spent working on the task. Successful reasoners were slightly faster at completing the task but this difference was not statistically reliable, M_{Bayesian} = 124s, 95% CI = [98, 150], $M_{\text{incorrect}}$ = 155s, 95% CI = [116, 194], $t(35) = 1.34$, $p = .19$. Behavior durations were subjected to a 2-between × 4-within mixed analysis of variance (ANOVA). The between-subject factor was the final performance (incorrect vs. Bayesian) and the within-subject factor was the type of behavior (projection, marking, presentation change, or epistemic action; see Table 10 for a full definition). Results showed that participants spent different amounts of time engaging in the different types of behaviors coded, $F(3, 105) = 20.3$, $MSE = 648$, $p < .001$, $\eta^2_p = .37$. The average behavior duration did not reliably vary as a function of performance, M_{Bayesian} = 31.0s, 95% CI = [23.3, 38.7], $M_{\text{incorrect}}$ = 38.8s, 95% CI [29.9, 47.6], $F(1, 35) = 1.80$, $MSE = 302$, p

= .19, η^2_p = .05. However, there was a significant interaction between performance and behavior type, $F(3, 105) = 6.02$, $p < .001$, $\eta^2_p = .15$. Figure 5 illustrates this interaction. Fisher's LSD post hoc paired comparisons revealed that unsuccessful reasoners spent a significantly longer time marking the cards than they did on any other type of behavior. Then they engaged most in presentation change, and they spent the least amount of time engaging in epistemic actions and projection (i.e., thinking without touching or interacting with the cards). By contrast, successful reasoners spent significantly more time engaging in projection, marking, and presentation change than epistemic actions.

Finally, we subjected the proportion of time people spent on each type of behavior out of the total time they spent solving the task to the same 2-between × 4-within mixed ANOVA. As before, there was no reliable difference based on performance. On average, both successful and unsuccessful participants spent 25% of their time engaging with each type of behavior, $F < 1$. Overall, however, participants' time was not distributed equally between different types of behavior, $F(3, 105) = 22.7$, $MSE = 0.03$, $p < .001$, $\eta^2_p = .39$. This apparent contradiction is better explained by the significant interaction between behavior types and performance, $F(3, 105) = 6.94$, $p < .001$, $\eta^2_p = .17$. Bonferroni-corrected post hoc independent *t*-tests comparisons confirmed that Bayesian reasoners spent a significantly greater proportion of their time changing the presentation of the layout compared to unsuccessful reasoners, $t(35) = 3.16$, $p < .001$, Cohen's $d = 1.07$ whereas unsuccessful reasoners spent a significantly greater proportion of their time marking the cards, $t(35) = 3.57$, $p < .001$, Cohen's $d = 1.21$ (see Fig. 6).

To summarize, results from Experiment 4 corroborated the main finding of Experiments 1 and 2, namely increasing the level of interactivity afforded by a Bayesian reasoning task (using

playing cards to represent possibilities in the problems) greatly improved statistical reasoning. Moreover, this improvement appeared to be independent of particular characteristics of the material used to enhance interactivity: a similar pattern of results was observed when the information about the hypothesis (*H* vs. *not-H*) and the data (*D* vs. *not-D*) was printed on opposite sides of the playing cards (Experiment 1) or side-by-side (Experiment 4). Likewise, performance was not significantly impaired in the absence of a one-to-one correspondence between the sets described in the problems and the sets made of playing cards. The behavioral analysis of participants' hands movements clarified the processes by which participants may use the cards to enact their thinking, namely by acting directly on the structure of the information layout, laying out cards from the pack, picking them up from the table and putting them down in a new location or sliding them around to rearrange them before announcing their solution. Unsuccessful solvers, by contrast, appeared much less active, often spending several seconds holding the cards in their hands, touching or pointing at cards without moving their hands, or nudging them ever so slightly but without making any significant transformation to the layout.

Taken together, these experiments show that the active physical manipulation of the statistical information in Bayesian tasks can transform statistical reasoning above and beyond the support offered by presenting the problem information in a frequency format. Admittedly, however, those experiments have not provided direct evidence for the claim that it is the manipulation of the cards *per se* that causes the increase of performance. One could argue that the cards merely offer an alternative way of representing the frequency counts using a particular, discrete, countable and iconic representation. Recent research has shown that providing a representation of the statistical information with discrete icons led to a significant increase in

Bayesian reasoning performance (Brase, 2009), although iconicity may not matter so much as the provision of a visual presentation of nested sets using countable dots (Sirota et al., 2014). In any case, it could be that the manipulation of the cards itself is irrelevant to the success rate observed in our experiments. Instead, it could be that the card manipulation resulted in a congenial external representation of the statistical data, which itself was responsible for the increase in performance. Alternatively, if the physical manipulation of cards did matter, it may be merely because it increased participants' engagement with the task rather than because it transformed their cognitive processing of the task.

Experiment 5

Experiment 5 was designed to test whether the manipulation of cards has a direct influence on statistical reasoning, above and beyond an incidental increase in participants' involvement with the task. Specifically, this experiment had two aims. First, it aimed to examine whether it was the physical manipulation of cards rather than the discrete and countable layout resulting from this manipulation, which improved statistical reasoning. Second, it aimed to examine whether the effect of the physical manipulation was mediated by participants' level of engagement with the task.

Method

Participants. A total of 40 Psychology Students (35 women, 5 men, mean age = 24 years, *SD* = 9.43) took part in the experiment in exchange for course credits. The data were collected individually in the Kingston Psychology Observation Laboratory. The experiment was conducted in English.

Design and procedure. Participants were invited to solve a series of three Bayesian reasoning problems presenting the statistical information using natural frequency statements. We used three scenarios adapted from Zhu and Gigerenzer (2006): the cat problem, the glasses problem, and the teeth problem. We used the frequency sets 4, 5, and 6 from Experiment 4 (see Table 3). Scenarios were rotated to produce two orders, which were randomly allocated to participants. Upon consenting to participate and to be filmed, participants were sat at a table. Half of the participants were asked to solve three Bayesian problems with the help of an associated pack of cards, using the same materials and procedure used in Experiment 4 (isomorphic samples). The remaining half of the participants were presented with a laminated A3 sheet picturing the sample of cards in a scatter (see Fig. 7 for an illustration of the Glasses problem). They were instructed to keep their hands and fingers still on the tabletop while they thought about the problem. To minimize cognitive load, similar cards—cards representing one of the four categories of information, e.g., all the Pr($D|H$) cards—were represented in close proximity on the sheet. Neither pen nor paper was provided. Instead, participants were instructed to ring a bell once they were ready to announce their answer, at which time the experimenter re-entered the room to record their answer. Following the completion of each problem, participants were asked to complete a "flow" experience questionnaire. Flow is conceptualized as a mental state where one is deeply absorbed in an activity, which balances one's skills and the challenge offered by the activity (Csikszentmihalyi, 1990, 1997). Following Shernoff, Csikszentmihalyi, Shneider, and Shernoff (2003), we designed a 9-item scale measuring five key dimensions of the experience of flow: engagement (measured through concentration, interest and enjoyment), challenge, skill, and attention (measured through distraction and focus) and state of mind

(measured through anxiety and relaxation). The item composition of the scale is presented in Table 12. Each item was rated on an 8-point scale ranging from 0 (definitely not) to 7 (definitely yes). After the third task was completed, participants filled in the numeracy scale developed by Lipkus et al.'s (2001), were thanked and debriefed. Thus, a 2 (physical manipulation) between-subject design was used, with flow and numeracy as continuous predictors.

Results and Discussion

Bayesian performance. Answers were classified as Bayesian using a strict outcome criterion as in previous experiments. The primary objective of this experiment was to examine whether the manipulation of cards itself, rather than the iconic external representation they provide, was responsible for the increase in performance observed. To answer this question, we tested for between-subject effects by regressing average performance over the three problems on three predictors: interactivity (contrast-coded -1 for absent and 1 for present), the mean deviation form of the numeracy score, and the product of these two variables. The descriptive statistics and intercorrelations for the model variables are presented in Table 13. The results of the regression analysis are presented in Table 14. The increase in interactivity offered by the cards was associated with a significant increase in performance; $M_{\text{Low\_interactivity}} = .52$, $SD = .35$, $M_{\text{High\_interactivity}} = .77$, $SD = .31$, $p = 0.0098$. Higher numeracy scores also resulted in better performance, $p = .002$. There was, however, no evidence that numeracy moderated the effect of interactivity on performance, $p = .384$.

Interactivity and flow. A secondary aim of this experiment was to examine whether the effect of interacting with cards on performance was mediated by participants' level of engagement with the task, using a 9-item scale to measure flow. To test this mediation

hypothesis, we adopted a bootstrapping approach (Preacher & Hayes, 2004) to assess the indirect effect of interactivity on performance through flow, using 10,000 bootstrap samples. Results confirmed a direct effect of interactivity on performance, $B_{direct} = .115$, $SE = .06$, $t(37) = 2.07$, $p = .045$. However, the 95% bias-corrected confidence interval for the size of the indirect effect included zero, thus showing no evidence that the manipulation of cards had a positive effect on performance through an increase in flow or engagement with the task, $B_{indirect} = .010$, bootstrap $SE = .017$, 95% bias-corrected bootstrap CI [- .017, .053].

Altogether, these results confirmed it is the manipulation of cards, rather than their physical representation, which is responsible for the increase of performance observed. In the absence of physical manipulation, the rate of Bayesian performance was similar to that observed by Brase (2009; i.e., circa 50%). Physical manipulation, by comparison, resulted in significantly higher success rates (circa 75%), demonstrating the sizeable impact of physical manipulation over physical representation in this group of participants. This effect was not mediated by an amplified experience of flow while manipulating the cards. Taken together, these results show that physical manipulation does not merely facilitate information representation, or increase reasoners' engagement with the task; it has a direct and sizeable effect on statistical reasoning itself. And although those with higher numeracy skills performed better than those with lower numeracy skills, there was no evidence to show that the positive effect of manipulating cards while thinking through the problem was moderated by participants' numeracy skills.

General Discussion

Informed by a systemic approach to cognition (e.g., Hutchins, 2001; Villejoubert, & Vallée-Tourangeau, 2011; Vallée-Tourangeau & Vallée-Tourangeau, 2014), we surmised that

increasing the manipulability of the problem information in Bayesian reasoning tasks would enable participants to solve these tasks successfully without training. Our results clearly supported this prediction, not only for tasks using natural frequency statements (Experiments 1, 4 and 5), but also for tasks using single-event probability statements (Experiment 2). Importantly, we provided direct evidence that the improved performance was caused by the physical manipulation of the material apparatus, and not by the final, static, material presentation of the cards layout (Experiment 5). A secondary hypothesis was that the positive effect of an increased manipulability of the task information would be moderated by individual differences in numeracy skills. This was not confirmed by our results. Although numeracy was a significant predictor of Bayesian performance, it did not moderate the effect of the increased manipulability of the task on performance.

The generalizability and the reliability of the effect of increasing manipulability on performance was apparent in Experiment 4, where high levels of performance were observed even in the absence of a perfect match between the statistical data in the problem statement and the number of cards in the deck provided to participants, and even when the cards presented information about data and hypothesis side by side on one front rather than on two opposite sides. Experiment 3 demonstrated that the explicit mention of all probabilities in a traditional pen-and-pencil Bayesian task had no significant impact on participants' performance. Moreover, the systematic analysis of the videographic evidence in Experiment 4 showed that successful performance was underpinned by an active manipulation of the cards leading to a change in the presentation layout whereas unsuccessful performance was characterized by a lesser degree of manipulation and an increase in marking behavior such as touching or holding cards without

moving them. Taken together, these results support the view that the physical actions afforded by the use of cards to present the problem information (rather the explicit description of all elements in the sample space) promotes correct statistical reasoning.

Systemic thinking: A dual-flow model of cognition

Whereas the positive impact of higher levels of information manipulability on performance could be anticipated from the distributed cognition perspective, the causal pathways through which manipulability actually lead to an increase in performance remains only loosely accounted for. Figure 8 provides an illustration of the classical information-processing model adapted from Baddeley's (2012) working memory model. This model artificially sequesters cognition in a series of singular Input-Processing-Output events and provides a rather idealistic view of human cognition as a planned information processing process: we see, we think, we act. From this perspective, errors of performance are assumed to arise from a breakdown in this mental subroutine and are attributed to a faulty mental representation, a shortage of individual knowledge, cognitive resources, or motivation (e.g., Kahneman, 2003; Darlow & Sloman, 2010).

This procrustean model of cognition cannot account for the results we report in the present study, which suggest instead that thinking can be shaped by action: we see, we act, we think. For example, motor actions enhance memory (Cook, Yip, & Goldin-Meadow, 2010), mental arithmetic (Carlson et al., 2007; Vallée-Tourangeau, 2013), and insight (Vallée-Tourangeau, in press). Thinking, reasoning, and deciding, we contend, would be better modeled by a dual-flow model of processing, where cognition arises from one of two distinct processing pathways: A *deductive* loop where the next action, response, or behavior is deduced from the cognitive processing of a mental representation and an *inductive* loop where the next action, response or

behavior is induced from the affordances offered by the immediate environment Figure 9 illustrates this alternative processing model, which we call the Systemic Thinking model (SysTM).

Thus, when cognition flows through the deductive processing loop, the perception of a stimulus (e.g., the printed text of a Bayesian reasoning task) contributes to the shaping of a mental representation (e.g., a sample space), which is processed internally, inside the head. The stimulus shapes the representation and different representations may afford different processing. For example, a Bayesian task presenting information using single-event probability statements may be represented as a sample space including multiple sets and different representations may afford different cognitive algorithms (e.g., see Villejoubert & Mandel, 2002, Fig. 1, p. 173). Conversely, a Bayesian task presenting information using natural frequency statements, for example, may shape a nested set representation, which in turn will afford the application of cognitive operations that will result in the production of a normative answer (Gigerenzer & Hoffrage, 1995; Sirota et al., 2015; Sloman et al., 2003).

But cognition does not need to always arise from such a deductive processing loop. In other instances, the material presentation of a stimulus may elicit the direct perception of an affordance (i.e., an action possibility) and give rise to the physical processing of the material presentation controlled by a motor executive (as opposed to a central executive, which orchestrates cognitive processing). As Figure 9 illustrates, such an inductive loop involves the procedural long-term memory storage as well as an "affordance pool" (allowing the direct perception of action possibilities), which sits alongside the visuo-spatial sketchpad and the phonological loop of Baddeley's working memory model. In other words, a stimulus may serves as an online guide for

action, and as such, does not require its mental recognition and classification (Baber et al., 2014; Gibson, 1979/1986; Greeno, 1994; Norman, 2002; Withagen & Chemero, 2012). For instance, a pack of cards affords epistemic actions (e.g., sampling through a pack of cards, see Table 10) that do not require mental representations or purposeful cognitive planning to take place. Not all actions need to result from an *a priori* mental plan, they can arise as people "follow materials" (Ingold, 2009) in a spatio-temporal trajectory where thinking is shaped by the intertwining of people's internal resources and the external artifacts at their disposal (Vallée-Tourangeau & Vallée-Tourangeau, 2014). While engaged in such an inductive processing loop, the reasoner's activity is not dictated by a mental representation of the task; instead it forms the cognitive substrate from which the mental representation can emerge. While they take place, actions that arise directly from perceptual affordances may shape cognitive processing of the information sampled without the need for an intermediate mental representation as people process information through their actions (e.g., by restructuring the layout of cards). When cognition flows through such an inductive loop, what people do informs what they think, as illustrated by the behavioral results observed in Experiment 4: participants who *actively* rearranged the cards were also more likely to correctly solve the problems.

Whereas the classical information processing model incorporate the bi-directional nature of information flow (with a top-down flow from memory storage to perception and a bottom-up flow from stimulus to perception), the concepts of inductive and deductive processing loops are unique to SysTM. In addition, the systemic thinking model implies that thinking and reasoning may not always follow a unique linear path that is either deductive (perception → mental representation → cognitive processing → physical processing) or inductive (perception →

physical processing) but may also loop locally during thinking. For example, a particular card layout may give rise to an epistemic action that leads to a reconfiguration of the layout, and this new configuration may itself make a new affordance salient (perception ⇄ physical processing) or inform cognitive processing (perception ⇄ cognitive processing). This example illustrates how manipulable physical apparatuses that allow participants to engage in epistemic and restructuring actions may augment and transform their capacity to process the task at hand through a dynamic spatio-temporal trajectory (see also Vallée-Tourangeau & Vallée-Tourangeau, 2014).

Implications and Future Research

Bayesian reasoning. The Systemic Thinking Model has important theoretical implications for the study of Bayesian reasoning in particular, and our conception of human cognition in general. As far as Bayesian reasoning is concerned, neither the ecological rationality account nor the nested-set account can explain the findings reported here because both these accounts are informed by the same basic assumption: namely that cognition emerges from information processing that is carried out inside the head, on an initial mental representation which mirrors a static perceptual input. In other words, these two accounts, akin to most mainstream cognitivist accounts, presuppose that the main function of people's neural activity is to process information, and that, to realize this function, the nervous system begins by mimicking the properties of the environment in the form of more or less accurate mental constructions. These accounts disagree on the properties of the mental representations needed to enable effective cognitive processing. The ecological-rationality account argues that the mental representation must include individuated objects with a natural-sampling structure because cognitive processing is

constrained by evolutionary designed frequency-coding mechanisms (e.g., Brase, 2002; Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995). The nested-set account argues that the natural-sampling structure constraint is too stringent: any mental representation that highlights the nested-set structure of the task information will afford successful cognitive processing and the production of a Bayesian response (e.g., Girotto & Gonzalez, 2001; Sirota et al., 2015; Sloman et al. 2003). But neither account anticipates the causal role of physical action on cognitive processing evidenced in the series of experiments we report here.

Natural frequency formats and nested set relationships play an undeniable part in facilitating Bayesian reasoning but only a systemic perspective can ultimately enable a comprehensive understanding of how environments can be better designed to foster accurate statistical reasoning. Our experiments show that enabling physical actions promoted performance beyond and above information format; this highlights the need to transcend debates about how the mind may best represent the world (as in current debates between proponents of the nested-set and the ecological-rationality accounts; see, e.g., Barbey & Sloman, 2007) and instead focus on gaining a better understanding how the properties of the mind, the body, and the environment complement each other to produce the results observed (Vallée-Tourangeau & Vallée-Tourangeau, 2014).Admittedly, approaching the study of cognition through the lens of the classical information-processing model does not necessarily entails a commitment to "methodological solipsism" (Fodor, 1980) where mental processing is conceived in isolation from the physical world within which it takes place. The impact of the environment on internal representations and mental processing has long been featured in classical information-processing models. The ecological rationality approach to cognition indeed argues that we need to understand how the

mind exploits its environment to understand its cognitive machinery (Brighton & Todd, 2001). These accounts, however, remain subordinate to a dualist conception of the mind and the environment. Cognition classically conceived may be assumed to have evolved its computational mechanisms in symbiosis with nature, but it remains viewed as emerging from a "self-actional" brain, which ultimately represents and computes information "offline", through a somewhat linear succession of mental states and mental processes. That people are able to think and solve problems "offline" (e.g., see Clark, 2010; Wilson, 2002) does not necessarily imply that this is how they all think typically or that this is how they all think best. In other words, the classical information-processing model stands as a blinker for the constitutive role of physical actions in the genesis of cognition, which is, in turn, traditionally studied in ecosystems that severely limits opportunities to act upon the information to be processed—that is, experimental procedures wherein interactivity is reduced or eliminated.

The systemic thinking model does not merely seek to fuse the nature of the environmental input or the top-down and bottom-up nature of information flow in such an "offline" information-processing model. It calls, instead, for a renewed conception of higher cognition that incorporates the succession of events taking place online, not only inside a cognizing agent's brain, but also outside, in the form of physical actions performed in her immediate environment. This presupposes that optimal results (e.g., producing a Bayesian inference) are unlikely to be a function of the quality of the reproduction of the environment characteristics in a representation or a model; instead they are a function of the *degree of fit* between (i) the cognitive processes and abilities possessed by a given individual, (ii) the physical actions that are enacted by the body and (iii) the affordances of the environment (e.g., see Anderson, 2014; Järvilehto, 1998).

Implications for the study of thinking and reasoning. The role of affordances and inductive cognitive pathways on thinking, reasoning and decision-making is ubiquitous. Material artefacts such as cards, paper and pencil and other tools form a constitutive part of individuals' ability to think and yet their role often goes unnoticed. For example, our results suggest that, in all likelihood, interactivity with the task material in Cosmides and Tooby's (1996) "active pictorial task" contributed to the record 92% success rate they observed since participants were instructed to circle parts of the visual representation of frequencies to represent base rates and false-alarm rates in the written task statement before providing an answer. But the scaffolding role of artefacts in thinking extends beyond Bayesian reasoning. M. Oaksford (personal communication, January 4, 2012) noted that data selection behaviour in a hypothesis-testing task was better aligned with the predictions of the Optimal Data Selection model (Oaksford & Chater, 1994) when the procedure involved differently sized stacks of cards to reflect varying probability manipulations (see Oaksford, Chater, Grainger & Larking, 1997; Oaskford, Chater, & Grainger, 1999, Experiment 4; Oaksford & Wakefield, 2003; Oaksford & Moussakowski, 2004, Experiment 2). In other words, providing individuals with material artefacts can enhance their hypothesis testing performance. In like vein, in a recent paper examining the different cognitive processes that underpin insight in problem-solving, Fleck and Weisberg (2013) reported data from five test problems, three of which made use of manipulable apparatuses while the remaining two did not. They remarked that restructuring, defined as a change in a reasoner's representation of the problem, occurred more often in some problems than others. What the authors failed to notice, however, was that restructuring and ultimately insight were associated with interactivity: participants were more likely to engage in restructuring when the material

presentation of the task afforded physical processing (see also Steffensen, Vallée-Tourangeau, & Vallée-Tourangeau, under review; Vallée-Tourangeau, in press; Weller et al., 2011).

The Systemic Thinking Model offers substantial heuristic value for guiding future research as it serves as a springboard for testing the situated parameters that may affect the relative importance of the deductive or inductive pathways in thinking and decision-making. These situated parameters could reflect the characteristics of the reasoner (e.g., does higher working memory capacity or domain relevant expertise moderate the effect of interactivity on performance?), the situation (e.g., does cognitive load and task difficulty encourage or discourage physical processing?), and the environment (e.g., what affordances promote or hinder optimal reasoning?).

Conclusions

When the task material affords the restructuration of the problem data in the world, rather than in the head, performance leaps up. This is both unsurprising and far-reaching. It is unsurprising because, as cognitive psychologists, we "know" that props can support cognitive activities. Teachers use props to support mathematical thinking in young children (e.g., Martignon & Krauss, 2009). Neuropsychologists use props to assess memory in the elderly (e.g., Anderson-Hanley, Miele, & Dunnam, 2012). Researchers, educators and rehabilitators conceive props as an aid to those who have not yet fully developed or have lost some of their cognitive potential. In that respect, the potential of the SysTM perspective goes beyond the main result that performance improves with props. What our findings show is that performance improves with interactivity; that is, when a participant's nascent Bayesian solution develops through the dynamic coupling with a malleable physical presentation of the problem. Thus, the Systemic

Thinking model does not simply make predictions about the effect of making situations more concrete, the presence or absence of props, or indeed about the importance of the environment in the classically mapped interaction between so-called top-down and bottom-up processes. Instead, it highlights the need to account for how a problem's solution emerges through the "interweaving" of physical processing and cognitive processing. This perspective casts aside the ontological debate sparked by the extended mind hypothesis (e.g., Clark & Chalmers, 1998) by affirming that the important issue is no longer where cognitive processing begins and where it ends, but rather *how* cognition emerges from the interactions of brain activity, motor actions, and artefacts (see also Vallée-Tourangeau & Vallée-Tourangeau, 2014).

Most cognitive psychologists have long assumed healthy adults ought to be able to manipulate ideas in their head. Piaget's (1928) proposition that once individuals have reached the formal operational stage they no longer depend on concrete manipulations in thinking, remains by and large unchallenged. The present research, by contrast, shows that the complex reasoning of healthy adults can also be transformed by having the opportunity to manipulate physical constituents of the problems encountered. In that sense, our findings are far-reaching: they call us to consider the possibility that the error was not only Descartes'; it might have been Piaget's also.

References

Anderson, M. L. (2014). *After phrenology: Neural reuse and the interactive brain*. Cambridge, MA: MIT Press.

Anderson-Hanley, C., Miele, A. S., & Dunnam, M. (2013). The Fuld Object-Memory Evaluation: Development and validation of an alternate Form. *Applied Neuropsychology*, *20*, 1–6. doi:10.1080/09084282.2012.670156

Baber, C., Parekh, M., & Cengiz, T. G. (2014). Tool use as distributed cognition: how tools help, hinder and define manual skill. *Frontiers in psychology, 5* :116. doi: 10.3389/fpsyg.2014.00116

Baddeley, A. (2012). Working Memory: Theories, Models, and Controversies. *Annual Review of Psychology*, *63*, 1–29. doi:10.1146/annurev-psych-120710-100422

Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences*. Cambridge, NY: Cambridge University Press.

Bar-Hillel, M. (1983). The base rate fallacy controversy. In R. W. Scholz (Ed.). *Decision making under uncertainty: Cognitive decision research, social interaction, development and epistemology*. (pp. 39–61). Amsterdam: Elsevier Science.

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences, 30*, 241–297. doi:10.1017/S0140525X07001653

Brase, G. L. (2002). Which statistical formats facilitate what decisions? The perception and influence of different statistical information formats. *Journal of Behavioral Decision Making*, *15*, 381–401.

Brase, G. L. (2009). Pictorial Representations in Statistical Reasoning. *Applied Cognitive Psychology*, *23*, 369–381. doi:10.1002/acp.1460

Brase, G. L., Fiddick, L., & Harries, C. (2006). Participant recruitment methods and statistical reasoning performance. *Quarterly Journal of Experimental Psychology, 59*, 965–76. doi:10.1080/02724980543000132

Brighton, H., & Todd, P. M. (2009). Situating rationality: ecologically rational decision making with simple heuristics. In P. Robbins & M. Aydede (Eds.), *The Cambridge Handbook of Situated Cognition* (pp. 322–346). New York, NY: Cambridge University Press.

Carlson, R. A., Avraamides, M. N., Cary, M., & Strasberg, S. (2007). What do the hands externalize in simple arithmetic? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 747. doi:10.1037/0278-7393.33.4.747

Chapman, G. B., & Liu, J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgment and Decision Making, 4*, 34–40.

Clark, A. (2010). Material surrogacy and the supernatural: Reflections on the role of artefacts in 'off-line' cognition. In L. Malafouris and C. Renfrew (Eds.), *The cognitive life of things: Recasting the boundaries of the mind* (pp. 23-28). Cambridge: McDonald Institute for Archaeological Research.

Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis, 58*, 7–19.

Cook, S. W., Yip, T. K., and Goldin-Meadow, S. (2010). Gesturing makes memories that last. *Journal of Memory and Language, 63*, 465–475.

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.

Csikszentmihalyi, M. (1991). *Flow: the psychology of optimal experience*. New York: Harper Perennial.

Csikszentmihalyi, M. (1997). *Finding flow: the psychology of engagement with everyday life*. New York: Basic Books.

Darlow, A. L., & Sloman, S. A. (2010). Two systems of reasoning: Architecture and relation to emotion. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 382–392. doi:10.1002/wcs.34

Fioratou, E., & Cowley, S. (2009). Insightful thinking: cognitive dynamics and material artifacts. *Pragmatics & Cognition, 17*, 549–572.

Fleck, J. I., & Weisberg, R. W. (2013). Insight versus analysis: Evidence for diverse methods in problem solving. *Journal of Cognitive Psychology*, *25*, 436–463. doi:10.1080/20445911.2013.779248

Fodor, J. A. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences*, *3*, 63–73. doi:10.1017/S0140525X00001771

Gibson, J. J. (1986). *The ecological approach to visual perception*. Hillsdale, NJ: Erlbaum. (original work published 1979).

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684–704. doi:10.1037/0033-295X.102.4.684

Gigerenzer, G., & Hoffrage U. (2007). The role of representation in Bayesian reasoning: correcting common misconceptions [Commentary on Barbey and Sloman]. *Behavioral and Brain Sciences*, *30*, 264-267.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge, UK: Cambridge University Press.

Girotto, V., & Gonzalez, M. (2001). Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition, 78*, 247–276.

Greeno, J. G. (1994). Gibson's affordances. *Psychological Review, 101*, 336–342. doi:10.1037/0033-295X.101.2.336.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*, 534–9. doi:10.1111/j.0956-7976.2004.00715.x

Hill, W. T., & Brase, G. L. (2012). When and for whom do frequencies facilitate performance? On the role of numerical literacy. *The Quarterly Journal of Experimental Psychology*, *65,* 2343–2368. doi:10.1080/17470218.2012.687004

Hoffrage U., Gigerenzer, G., Krauss S., & Martignon L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition*, *84*, 343-352.

Hurley, K., Marshall, J., Hogan, K., & Wells, R. (2012). A comparison of productivity and physical demands during parcel delivery using a standard and a prototype electric courier truck. *International Journal of Industrial Ergonomics*, *42*, 384–391. doi:10.1016/j.ergon.2012.04.003

Hutchins, E. (1995). How a cockpit remembers its speeds. *Cognitive Science, 19*, 265–288. doi:10.1207/s15516709cog1903_1

Hutchins, E. (2001). Cognition, Distributed. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 2068–2072). Oxford: Pergamon. doi:10.1016/B0-08-043076-7/01636-3.

Hutchins, E. (2010). Cognitive ecology. *Topics in Cognitive Science, 2*, 705-715. doi: 10.1111/j.1756-8765.2010.01089.x

Ingold, T. (2009). The textility of making. *Cambridge Journal of Economics, 34,* 91-102. doi: 10.1093/cje/bep042

Järvilehto, T. (1998). The theory of the organism-environment system: I. Description of the theory. *Integrative Physiological and Behavioral Science*, *33*, 321–334.

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2009). *Data analysis: a model comparison approach* (2nd ed.). New York ; Hove: Routledge.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American Psychologist*, *58*, 697–720. doi:10.1037/0003-066X.58.9.697

Kirsh, D. (1995). Complementary strategies: Why we use our hands when we think. In J. D. Moore and J. F. Lehman (Eds.) *Proceedings of the Seventheenth Annual Conference of the Cognitive Science Society* (pp. 212–217). Mahwah, NJ: Lawrence Erlbaum Associates.

Kirsh, D. (2009). Problem solving in situated cognition. In P. Robbins & M. Aydede (Eds.), *The Cambridge handbook of situated cognition* (pp. 264–306). Cambridge: Cambridge University Press.

Kirsh, D. (2013). Thinking with external representations. In S. J. Cowley and F. Vallée-Tourangeau (Eds.), *Cognition beyond the brain: Computation, interactivity and human artifice* (pp. 171-194). London: Springer-Verlag.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, *18*, 513–549. doi:10.1207/s15516709cog1804_1

Kleiter, G. (1994). Natural sampling: rationality without base rates. In G. H. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 375–388). New York, NY: Springer.

Koehler, J. J. (1996). The Base rate fallacy reconsidered: Normative, descriptive and methodological challenges. *Behavioral and Brain Sciences, 19*, 1–17.

Kurzenhäuser, S., & Hoffrage, U. (2002). Teaching Bayesian reasoning: An evaluation of a classroom tutorial for medical students. *Medical Teacher, 24*, 516–21. doi:10.1080/0142159021000012540.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21*, 37–44. doi:10.1177/0272989X0102100105.

Macchi, L. (2000). Partitive formulation of information in probabilistic problems: Beyond heuristics and frequency format explanations. *Organizational Behavior and Human Decision Processes, 82*, 217–236.

Martignon, L., & Krauss, S. (2009). Hands-on activities for fourth graders: A tool box for decision-making and reckoning with risk. *International Electronic Journal of Mathematics Education*, *4*, 227–258.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, *113*, 181–190.

McCloy, R., Beaman, C. P., Morgan, B., & Speed, R. (2007). Training conditional and cumulative risk judgements: the role of frequencies, problem-structure and einstellung. *Applied Cognitive Psychology, 21*, 325–344. doi: 10.1002/acp.1273.

Norman, J. (2002). Two visual systems and two theories of perception: An attempt to reconcile the constructivist and ecological approaches. *Behavioral and Brain Sciences*, *25*, 73–96.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101*, 608-631.

Oakford, M., Chater, N., & Grainger, B. (1999). Probabilistic effects in data selection. *Thinking and Reasoning, 5*, 193–243.

Oaksford, M., Chater, N., Grainger, B., & Larkin, J. (1997). Optimal data selection in the reduced array selection task (RAST). *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *23*, 441-458.

Oaksford, M., & Moussakowski, M. (2004). Negations and natural sampling in data selection: Ecological versus heuristic explanations of matching bias. *Memory & Cognition, 32,* 570–581.

Oaksford, M., & Wakefield, M. (2003). Data selection and natural sampling: Probabilities do matter. *Memory & Cognition, 31,* 143–154.

Peters, E., Vastfjall, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, *17*, 407–413. doi:10.1111/j.1467-9280.2006.01720.x

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability model inference task. *Journal of Experimental Psychology, 72*, 346–354.

Piaget, J. (1928). La causalité chez l'enfant. *British Journal of Psychology*, *18*, 276–301. doi:10.1111/j.2044-8295.1928.tb00466.x

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, *36*, 717–731.

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135*, 943–73. doi:10.1037/a0017327.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General, 130*, 380–400.

Shernoff, D. J., Csikszentmihalyi, M., Shneider, B., & Shernoff, E. S. (2003). Student engagement in high school classrooms from the perspective of flow theory. *School Psychology Quarterly*, *18*, 158–176. doi:10.1521/scpq.18.2.158.21860

Sirota, M., & Juanchich, M. (2011). Role of numeracy and cognitive reflection in Bayesian reasoning with natural fequencies. *Studia Psychologica, 53*, 151–161.

Sirota, M., Kostovičová, L., & Juanchich, M. (2014). The effect of iconicity of visual displays on statistical reasoning: evidence in favor of the null hypothesis. *Psychonomic Bulletin & Review*, *21*(4), 961–968. doi:10.3758/s13423-013-0555-4

Sirota, M., Kostovičová, L., & Vallée-Tourangeau, F. (2015). How to train your Bayesian: A problem-representation-transfer rather than a format-representation-shift explains training

effects. *Quarterly Journal of Experimental Psychology, 68*, 1-9. doi:10.1080/17470218.2014.972420

Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes,* 91, 296–309. doi: 10.1016/S0749-5978(03)00021-9.

Stanovich, K. E., & West, R. F. (1998). Who uses base rates and P(D/H)? An analysis of individual differences. *Memory & Cognition, 26*, 161–179.

Steffensen, S. V., Vallée-Tourangeau, F., & Vallée-Tourangeau, G. (under review). Cognitive events in a problem-solving task: Qualitative methods for investigating interactivity in the 17 animals problem

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

Vallée-Tourangeau, F. (in press). Insight, interactivity and materiality. *Pragmatics and Cognition*.

Vallée-Tourangeau, F. (2013). Interactivity, efficiency, and individual differences in mental arithmetic. *Experimental Psychology, 60*, 302-311. doi: 10.1027/1618-3169/a000200

Vallée-Tourangeau, F., Euden, G., & Hearn, V. (2011). Einstellung defused: Interactivity and mental set. *The Quarterly Journal of Experimental Psychology*, *64*, 1889–1895. doi:10.1080/17470218.2011.605151

Vallée-Tourangeau, F., & Payton, T. (2008). Graphical representation fosters discovery in the 2-4-6 task. *Quarterly Journal of Experimental Psychology, 61*, 625–640.

Vallée-Tourangeau, F., Payton, T., & Murphy, R. A. (2008). The impact of presentation format on causal inferences. *European Journal of Cognitive Psychology, 20*, 177–194.

Vallée-Tourangeau, F., & Wrightman, M. (2010). Interactive skills and individual differences in a word production task. *AI & Society*, *25*, 433–439. doi:10.1007/s00146-010-0270-x

Vallée-Tourangeau, G., & Vallée-Tourangeau, F. (2014). The Spatio-temporal dynamics of systemic thinking. *Cybernetics and Human Knowing, 21*, 113-127.

Vallée-Tourangeau, F., & Villejoubert, G. (2013). Naturalising problem solving. In S. J. Cowley, & F. Vallée-Tourangeau (Eds.), *Cognition beyond the brain: Interactivity, cognition and human artifice* (pp. 241-254). Dordrecht: Springer.

Villejoubert, G. (2007). Les aides graphiques peuvent-elles nous aider à mieux raisonner avec les probabilités ? [Can graphical aids help us better reason with probabilities?] *Annales de la Fondation Fyssen, 22*, 32–43.

Villejoubert, G., & Mandel, D. R. (2002). The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. *Memory & Cognition, 30*, 171–178. doi:10.3758/BF03195278.

Villejoubert, G., & Vallée-Tourangeau, F. (2011) Constructing preferences in the physical world: a distributed-cognition perspective on preferences and risky choices. *Frontiers in Cognition, 2*, 302. doi:10.3389/fpsyg.2011.00302

Vogel, M., Monesson, A., & Scott, L. S. (2012). Building biases in infancy: the influence of race on face and voice emotion matching: Influence of race on face and voice emotion matching. *Developmental Science*, *15*, 359–372. doi:10.1111/j.1467-7687.2012.01138.x

Weller, A., Villejoubert, G., & Vallée-Tourangeau, F. (2011). Interactive insight problem solving. *Thinking & Reasoning, 17*, 424–439. doi:10.1080/13546783.2011.629081.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*, 625–636.

Wilson, R. A., & Clark, A. (2009). How to situate cognition: Letting nature take its course. In P. Robbins & M. Aydede (Eds.), *The Cambridge handbook of situated cognition* (pp. 55–77). Cambridge: Cambridge University Press.

Withagen, R., & Chemero, A. (2012). Affordances and classification: On the significance of a sidebar in James Gibson's last book. *Philosophical Psychology*, *25*, 521–537. doi:10.1080/09515089.2011.579424

Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: Frequency or nested Sets? *Experimental Psychology, 50*, 97–106. doi:10.1026//1618-3169.50.2.97

Zhu, L., & Gigerenzer, G. (2006). Children can solve Bayesian problems: the role of representation in mental computation. *Cognition, 98*, 287–308. doi:10.1016/j.cognition.2004.12.003

Zhou, Y., Cameron, E., Forbes, G., & Humphris, G. (2012). Development of a novel coding scheme (SABICS) to record nurse–child interactive behaviours in a community dental preventive intervention. *Patient Education and Counseling*, *88*, 268–276. doi:10.1016/j.pec.2012.01.001

Footnotes

[1]We distinguish the way a probability *value* is expressed (i.e., as a percentage: 60%, a decimal: .60, or as a frequency ratio: 3 in 5) from the referent of a probability *statement*. A statement can refer to a unique outcome or to an outcome that is part of a set of identical outcomes. Thus, the statements: "the probability that a pupil watches too much TV is 60%" or "the chances that a pupil is watching too much TV are 3 in 5" are both single-event probability statements. In predicate logic, this would be represented as "$\exists x \mid P(x) \wedge \Pr[W(x)] = 60\%$" where P($x$) is the property "$x$ is a pupil" and W(x) is the property "x watches too much TV" or, in plain English, "There exists one element x such that x has the property "is a pupil" and x has a 60% probability to also have the property "watches too much TV". The same principle applies to a frequentist probability statement: using a frequency ratio to express the probability value is not what defines a probability statement as frequentist. Thus, the statement: "the probability that pupils watch too much TV is 60%" is a frequentist statement because its referent is a set of elements. It would be represented as $\boldsymbol{P} = \{x \mid P(x)\}$, $\boldsymbol{W} = \{x \mid W(x)\}$, $\forall x, x \in \boldsymbol{P} \longrightarrow \Pr(x \in \boldsymbol{W}) =$ 60% or in plain English, "for all elements x, if x belongs to the set of elements with the property "is a pupil", the probability that x also belong to the set of elements with the property "watches too much TV" is 60%". Natural frequency statements are particular exemplars of frequentist statements that restrict the format of the probability value to raw (i.e., non-normalized) frequency ratios, as they would be experienced through natural sampling.

[2]The Observer XT is a software application used to code and analyze observational data. It had been used in several areas of research, such as infant studies (Vogel, Monesson, & Scott, 2012), doctor-patient interaction studies (Zhou, Cameron, Forbes, & Humphris, 2012), or

ergonomics research (Hurley, Marshall, Hogan & Wells, 2012). It facilitates the development of coding manuals, the coding of video data, as well as the conduct of inter-rater reliability analyses.

Table 1

Sets of frequencies and conditional probabilities used in Experiments 1 and 2, respectively

	Set 1		Set 2		Set 3	
	Frequency	Probability	Frequency	Probability	Frequency	Probability
Base rate	12/20	60%	15/20	75%	4/20	20%
Hit rate	6/12	50%	9/15	60%	3/4	75%
False-alarm rate	2/8	25%	1/5	20%	8/16	50%
Bayesian answer	6/8	75%	9/10	90%	3/11	27%

Table 2

Means, Standard Deviations, and Intercorrelations for Numeracy and Bayesian Performance Variables in Experiment 1

Variable	*M*	*SD*	1	2
Predictors				
1. Interactivity (-1 = absent, 1 = present)			—	
2. Numeracy	8.61	1.85	-.03	—
Outcomes				
3. Mean Bayesian Performance (Y0)	0.62	0.40	.26*	.27*
4. Difference in Performance (Y1)	0.13	0.50	.22*	.03

Note. Y0 = $\frac{1}{3}\sum_{3}^{1} y_i$. Y1 = $y_3 - y_1$, where y_i is the performance on trial *i*.

*$p < .05$.

Table 3

Regression Analysis Summary for Interactivity, Numeracy, and Practice Predicting Bayesian Performance in Experiment 1

Variable	*B*	95% CI	ß	*t*	*p*
Outcome: Mean Bayesian Performance (Y0)					
Overall	0.62	[0.54, 0.7]	–	15.71	**< .001**
Interactivity	0.11	[0.03, 0.19]	0.27	2.71	**.008**
Numeracy	0.06	[0.01, 0.1]	0.27	2.66	**.009**
Interactivity × Numeracy	-0.02	[-0.07, 0.02]	-0.10	-1.04	.300
Outcome: Difference in Performance (Y1)					
Practice	0.132	[0.03, 0.24]	–	2.53	**.013**
Interactivity × Practice	0.112	[0.01, 0.22]	0.22	2.134	**.036**
Numeracy × Practice	0.009	[-0.05, 0.07]	0.03	0.322	.748
Interactivity × Numeracy × Practice	-0.02	[-0.07, 0.04]	-0.06	-0.60	.549

Note. R^2 for Mean Performance = .15. R^2 for Performance Difference = .05. CI = confidence interval for *B*.

Table 4

Means, Standard Deviations, and Intercorrelations for Numeracy and Bayesian Performance Variables in Experiment 2

Variable	*M*	*SD*	1	2
Predictors				
1. Interactivity (-1 = absent, 1 = present)			—	
2. Numeracy	8.86	1.83	.07	—
Outcomes				
3. Mean Bayesian Performance (Y0)	0.33	0.42	.57***	.26*
4. Difference in Performance (Y1)	0.04	0.33	.14	-.16

Note. Y0 = $\frac{1}{3}\sum_{3}^{1} y_i$. Y1 = $y_3 - y_1$, where y_i is the performance on trial *i*.
*$p < .05$. ***$p < .001$.

Table 5

Regression Analysis Summary for Interactivity, Numeracy, and Practice Predicting Bayesian Performance in Experiment 2

Variable	*B*	95% CI	ß	*t*	*p*
Outcome: Mean Bayesian Performance (Y0)					
Overall	0.33	[0.26, 0.4]	–	9.24	**< .001**
Interactivity	0.24	[0.17, 0.31]	0.56	6.67	**< .001**
Numeracy	0.05	[0.01, 0.09]	0.20	2.38	**.020**
Interactivity × Numeracy	0.03	[-0.01, 0.07]	0.14	1.68	.097
Outcome: Difference in Performance (Y1)					
Practice	0.043	[-0.03, 0.11]	–	1.246	.216
Interactivity × Practice	0.048	[-0.02, 0.12]	0.15	1.385	.170
Numeracy × Practice	-0.03	[-0.07, 0.01]	-0.17	-1.612	.111
Interactivity × Numeracy × Practice	0.008	[-0.03, 0.05]	0.05	0.43	.670

Note. R^2 for Mean Performance = .40. R^2 for Performance Difference = .05. CI = confidence interval for *B*.

Table 6

Sequences of Card Manipulations Observed on the First Problem Solved by Participants in the High-interactivity Condition of Experiment 2

Card manipulation (Experiment 2, High-interactivity condition, First problem)	Total
Sorts *H* and *not-H*; Among *H*, sorts *D* and *not-D*; Repeats among *not-H*; Selects *D*&*H* and *D*&*not-H*; Counts *D*; Counts *D*&*H*; Reports *D*&*H* out of *D*.	9
Sorts *D* and *not-D*; Counts *D*; Sorts *D*&*H* and *D*&*not-H*; Counts *D*&*H*; Reports *D*&*H* out of *D*.	7
Sorts *H* and *not-H*; Counts *D*&*H*; Reports *D*&*H* out of *Total*.	3
Does not use the cards.	2
Sorts *H* and *not-H*; Among *H*, sorts *D* and *not-D*; Repeats among *not-H*; Selects *D*&*H* and *D*&*not-H*; Counts *D*; Counts *D*&*H*; Reports *D* out of *Total*	2
Sorts *H* and *not-H*; Among *H*, sorts *D* and *not-D*; Repeats among *not-H*; Counts *D*&*H*; Counts *D*&*not-H*; Reports *D*&*H* - *D*&*not-H*.	1
Sorts *H* and *not-H*; Among *H*, sorts *D* and *not-D*; Repeats among *not-H*; Counts Total; Counts *H*; Reports *H* out of *Total*.	1
Sorts *H* and *not-H*; Counts *H*; Counts *D*&*H*; Report *D*&*H* out of *H*.	1
Sorts *H* and *not-H*; Counts *H*; Reports *H* out of *Total*.	1
Grand Total	27

Table 7

Sets of Frequencies Used in Experiment 4

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6
Base rate	4/20	14/20	6/20	6/20	8/20	12/20
Hit rate	3/4	7/14	4/6	2/6	7/8	8/12
False-alarm rate	6/16	2/6	6/8	7/14	6/12	2/8
Bayesian answer	3/9 (33%)	7/9 (78%)	4/10 (40%)	2/9 (22%)	7/13 (54%)	8/10 (80%)

Table 8

Means, Standard Deviations, and Intercorrelations for Numeracy and Bayesian Performance Variables in Experiment 4

Variable	*M*	*SD*	1
Predictor			
1. Numeracy (X)	8.20	2.14	—
Outcomes			
2. Mean performance (Y0)	0.58	0.38	.26
3. Difference in Performance by Sample type (Y1)	0.00	0.34	-.05
4. Difference in Performance by Practice (Y2)	0.23	0.38	-.25
5. Difference in Performance by Sample type and Practice (Y3)	-0.25	0.72	.17

Note. Y0 = $\frac{1}{6}\sum_{6}^{1} y_i$. Y1 = $\left(\frac{1}{3}\sum_{6}^{4} y_i\right) - \left(\frac{1}{3}\sum_{3}^{1} y_i\right)$. Y2 = $\frac{1}{2}(y_3 + y_6) - \frac{1}{2}(y_1 + y_4)$. Y3 = $(y_6 - y_4) - (y_3 - y_1)$, where y_i is the performance on trial *i*.

Table 9

Regression Analysis Summary for Numeracy, Type of Sample and Practice Predicting Bayesian Performance in Experiment 4

Variable	*B*	95% CI	ß	*t*	*p*
Outcome: Mean Bayesian Performance (Y0)					
Overall	0.58	[0.41, 0.76]	–	6.91	**< .001**
Numeracy	0.05	[-0.04, 0.13]	0.26	1.13	.272
Outcome: Difference in Performance by Sampling (Y1)					
Type of Sample	0.00	[-0.17, 0.17]	–	0.00	1.00
Type of Sample × Numeracy	-0.01	[-0.09, 0.07]	-0.05	-0.20	.841
Outcome: Difference in Performance by Practice (Y2)					
Practice	0.225	[0.05, 0.4]	–	2.67	**.016**
Practice × Numeracy	-0.05	[-0.13, 0.04]	-0.25	-1.11	.283
Outcome: Difference in Performance by Sampling and Practice (Y3)					
Type of Sample × Practice	-0.25	[-0.59, 0.09]	–	-1.54	.140
Type of Sample × Numeracy × Practice	0.057	[-0.11, 0.22]	0.17	0.738	.470

Note. R^2 for Y0 = .07. R^2 for Y1 = .00. R^2 for Y2 = .06. R^2 for Y3 = .03. CI = confidence interval for *B*.

Table 10

Coding Scheme Used To Analyze Video Recordings in Experiment 4.

Activities and actions	Definition
0. Projection	No actions on cards.
0.1. No action	Looks at the cards or the task statement but neither hand is touching, pointing or hovering above cards.
1. Marking	Actions on cards that have no obvious epistemic or perceptual impact.
1.1. Nudges cards	Moves one or more card(s) slightly (< 2cm) on the table without significantly changing its/their location.
1.2. Marks cards	Hand or finger(s) touches, points or hovers above one or more card(s) without moving or nudging it/them.
1.3. Holds cards	Holds one or more card(s) without putting it/them down for at least 2 seconds.
2. Presentation change	Actions on cards that change the perceptual layout.
2.1. Picks up / puts down / lays out card(s)	Transfers one or more card(s) from the table to the hand(s) or from the hand(s) to the table.
2.2. Transforms cards layout	Significantly transforms the way cards are arranged on the table by sliding one or more card(s) (> 2cm) to reorder it/them or move it/them to a completely different location.
3. Epistemic actions	Actions on cards that enable information processing.
3.1. Samples cards	Examines 3 or more cards in hands by flicking through them (> 2s).
3.2. Counts cards	Rapidly moves a hand or finger(s) from one card to the other over 3 cards or more.

Table 11

Example of Raw Observational Data Coded from the Video of a Successful Participant

Event	Event Start Time	Behavior
1	00:00:00	1.0 No action
2	00:00:02	2.2 Transforms cards layout
3	00:00:03	1.2. Marks cards
4	00:00:05	2.1 Picks up / puts down / lays out
5	00:00:06	1.1. Nudges cards
6	00:00:08	2.2 Transforms cards layout
7	00:00:16	1.1. Nudges cards
8	00:00:17	2.2 Transforms cards layout
9	00:00:24	1.1. Nudges cards
10	00:00:26	2.1 Picks up / puts down / lays out
11	00:00:36	1.1. Nudges cards
12	00:00:37	2.1 Picks up / puts down / lays out
13	00:00:42	1.1. Nudges cards
14	00:00:43	2.1 Picks up / puts down / lays out
15	00:00:44	1.1. Nudges cards
16	00:01:01	1.0 No action
17	00:01:03	1.2 Counts
18	00:01:09	1.0 No action
19	00:01:12	1.2 Counts
20	00:01:16	1.0 No action
21	00:01:23	1.2 Counts
22	00:01:26	1.0 No action
23	00:01:28	1.2. Marks cards
24	00:01:38	1.2 Counts
25	00:01:41	1.2. Marks cards
26	00:01:51	1.0 No action
27	00:01:52	Answers

Table 12

Item Composition of the Flow Scale

Items
Engagement
Were you able to concentrate well on the task?
Did you find the task interesting?
Did you enjoy what you were doing?
Challenge
Did you feel challenged by the task? (R)
Skill
Did you feel skilful while working on the task?
Attention
Did you feel focused while working on the task?
Did you feel distracted while working on the task? (R)
State of mind
Did you feel relaxed while working on the task?
Did you feel anxious while working on the task? (R)

Note. R = Reverse coded.

Table 13

Means, Standard Deviations, and Intercorrelations for Numeracy and Bayesian Performance Variables in Experiment 5

Variable	*M*	*SD*	1	2
Predictors				
1. Interactivity (-1 = absent, 1 = present)			—	
2. Numeracy	8.40	2.09	.00	—
Outcomes				
3. Mean Bayesian Performance (Y0)	0.64	0.35	.36*	.45**

Note. Y0 = $\frac{1}{3}\sum_{3}^{1} y_i$. Y1 = $y_3 - y_1$, where y_i is the performance on trial *i*.
*$p < .05$. **$p < .01$.

Table 14

Regression Analysis Summary for Interactivity and Numeracy Predicting Bayesian Performance in Experiment 5

Variable	*B*	95% CI	ß	*t*	*p*
Outcome: Mean Bayesian Performance (Y0)					
Overall	0.64	[0.55, 0.74]	–	13.77	**< .001**
Interactivity	0.13	[0.03, 0.22]	0.37	2.73	**< .010**
Numeracy	0.08	[0.03, 0.13]	0.48	3.44	**.002**
Interactivity × Numeracy	0.02	[-0.03, 0.07]	0.12	0.88	.384

Note. R^2 for Y0 = .35. CI = confidence interval for *B*.

Figure 1. Sample of cards used in the high-interactivity condition in Experiments 1 and 2.

Figure 2. Proportion of Bayesian answers with natural frequency statements as a function of interactivity level (low, paper-and-pencil only vs. high, paper-and-pencil with cards), and problem trial (first, second or third).

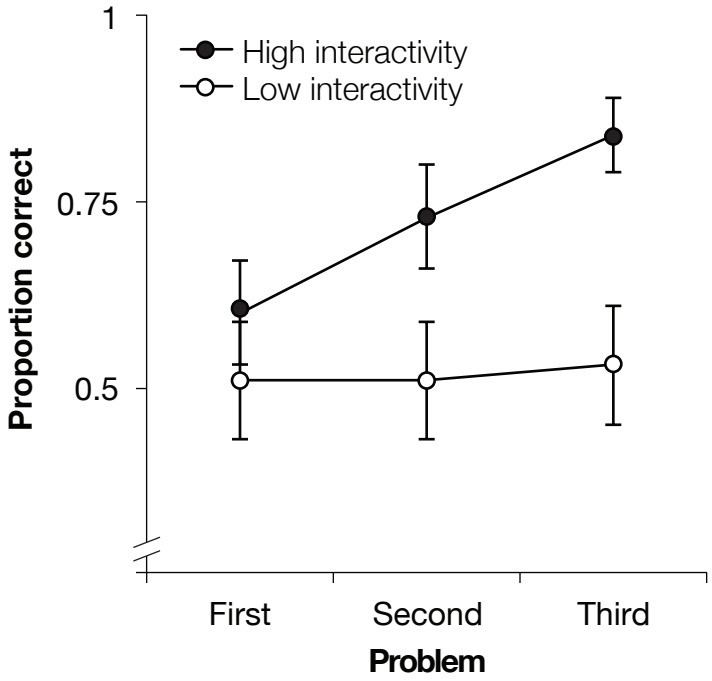


Figure 3. Sample of cards used in Experiment 4.

Figure 4. Snapshot of the video recording of a participant engaged in the task in Experiment 4.

Figure 5. Mean durations of the four categories of behavior coded. *$p < .05$, **$p < .01$, ***$p < .001$ denote significant mean differences based on Fisher's LSD post hoc paired comparisons.

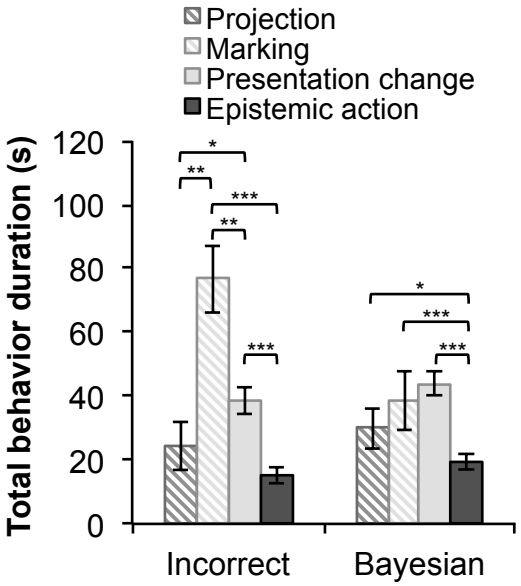


Figure 6. Mean percentage of time participant spent on each of the four categories of behavior coded. ***$p < .001$ denote significant mean differences based on Bonferroni-corrected post hoc independent t-tests comparisons. Proj = Projection; Mark = Marking; PrCh = Presentation Change; EpAc = Epistemic Action.

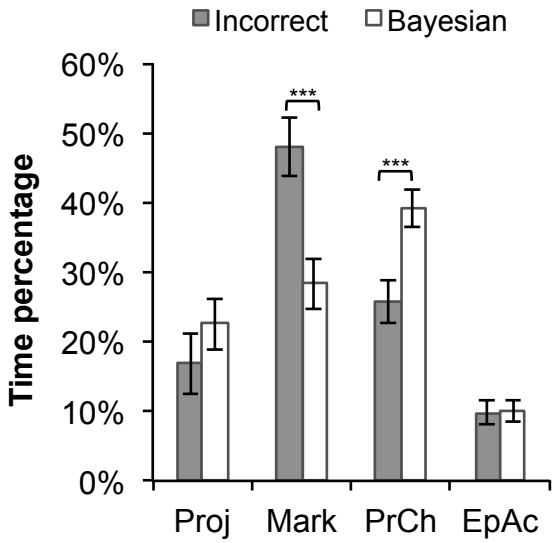


Figure 7. Illustration of the A3 laminated sheet presented to participants in the “hands still” condition (Interactivity absent, Glasses problem)

Figure 8. Schematic representation of the classical information-processing model (adapted from Baddeley, 2012).

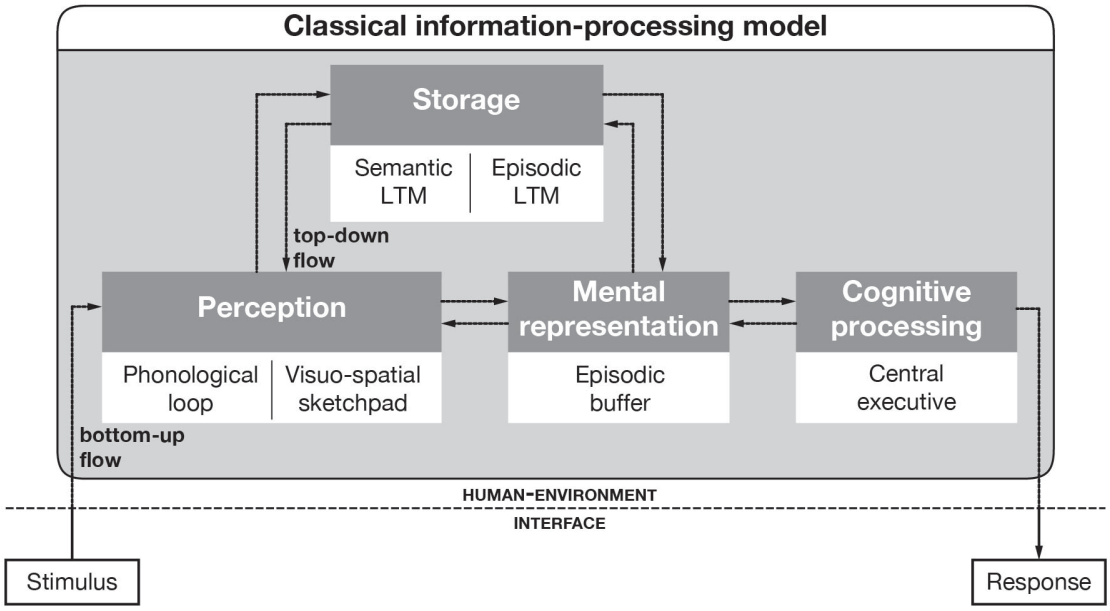
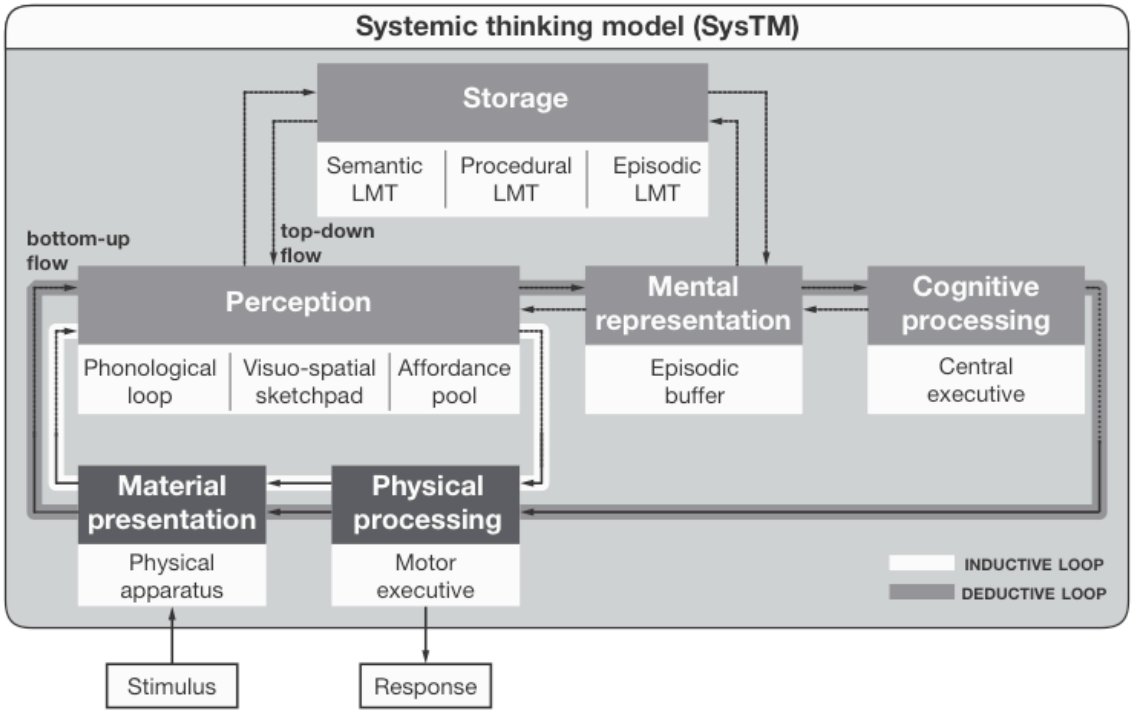


Figure 9. Schematic representation of the Systemic Thinking Model (SysTM).



Appendix

The 11 Items in the French Numeracy Scale adapted from Lipkus, Samsa, and Rimer (2001) with the Percentage of Participants Who Responded Correctly to Each Item in Experiments 1 and 2, as well as in Peters et al. (2006, Study 1) and in Lipkus et al.

Item	% correct			
	Exp. 1 (n = 90)	Exp. 2 (n = 90)	Peters et al. (n =100)	Lipkus et al. (n = 463)
1. *Imaginez qu'on lance 1 000 fois un dé non truqué à six faces. Sur les 1 000 lancers, combien de fois pensez-vous que le dé affichera un chiffre pair (2, 4 ou 6) ?* [Imagine that we rolled 1,000 times a fair, six-sided, die. Out of 1,000 rolls, how many times do you think the die would come up even (2, 4, or 6)?] Answer: 500 out of 1000.	58%	59%	61%	55%
2. *Dans la grande loterie de la Compagnie des Jeux, les chances de gagner un prix de 10 000 € sont de 1%. Si 1000 personnes achetaient chacune un ticket unique à la Compagnie des Jeux, quelle serait votre meilleure estimation du nombre de personnes gagnant un prix de 10 000 €?* [In the big lottery of the Games Company, the chance of winning a €10,000 prize is 1%. If 1,000 people each by a single ticket to the Games Company, what would be your best estimate of the number of people winning a €10,000 prize?] Answer: 10 out of 1000.	65%	74%	69%	60%
3. *Dans un Grand Jeu Concours d'une galerie commerciale, les chances de gagner une voiture sont de 1 sur 1 000. Quel est le pourcentage de tickets du Grand Jeu Concours qui gagnent une voiture ?* [In a Mall's Big Competition, the chance of winning a car is 1 in 1,000. What percent of tickets to the Big Competition win a car ?] Answer: 0.1%.	63%	65%	46%	21%
4. *Lequel des chiffres suivants représente le plus gros risque de contracter une maladie : 1 sur 100, 1 sur 1 000, 1 sur 10 ?* [Which of the following numbers represents the biggest risk of getting a disease? 1 in 100, 1 in 1,000, 1 in 10.] Answer: 1 in 10.	96%	96%	96%	78%
5. *Lequel des chiffres suivants représente le plus gros risque de contracter une maladie : 1%, 10%, 5% ?* [Which of the following numbers represents the biggest risk of getting a disease?] Answer: 10%.	96%	94%	94%	84%
6. *Si les risques de contracter une maladie pour une Personne A sont de 1% en dix ans et que les risques pour une Personne B sont le double de ceux de la Personne A, quels sont les risques pour la Personne B ?* [If Person A's risk of getting a disease is 1% in ten years, and person B's risk is double that of A's, what is B's risk?] Answer: 2%.	83%	87%	83%	91%
7. *Si les risques de contracter une maladie pour une Personne A sont de 1 sur 100 en dix ans et que les risques pour une Personne B sont le double de ceux pour la Personne A, quels sont les risques pour la Personne B ?* [If Person A's chance of getting a disease is 1 in 100 in ten years, and person B's risk is double that of A's, what is B's risk?] Answer: 2 in 100.	85%	89%	74%	87%

(continued)

Appendix (continued)

Item	% correct			
	Exp. 1 (n = 90)	Exp. 2 (n = 90)	Peters et al. (n =100)	Lipkus et al. (n = 463)
8A. *Si les chances de contracter une maladie sont de 10%, sur 100 personnes, combien de personnes sont supposées contracter la maladie ?* [If the chance of getting a disease is 10%, out of 100 people, how many would be expected toget the disease?] Answer: 10.	96%	96%	90%	81%
8B. *Si les chances de contracter une maladie sont de 10%, sur 1 000 personnes, combien de personnes sont supposées contracter la maladie ?* [If the chance of getting a disease is 10%, out of 1000 people, how many would be expected toget the disease?] Answer: 100.	75%	81%	84%	78%
9. *Si les chances de contracter une maladie sont de 20 sur 100, ceci équivaudrait à avoir ____% chances de contracter cette maladie.* [If the chance of getting a disease is 20 out of 100, this would be the same as having a ____% chance of getting the disease.] Answer: 20.	94%	94%	84%	70%
10. *Les chances de contracter une infection virale sont de .0005. Sur 10 000 personnes, combien sont supposées contracter la maladie ?* [The chance of getting a viral infection is .0005. Out of 10,000 people, about how many of them are expected to get infected?] Answer: 5.	52%	70%	56%	49%