Children's quantitative Bayesian inferences from natural frequencies and number of chances

Stefania Pighin [a,*], Vittorio Girotto [b], Katya Tentori [a]

[a] CIMeC, University of Trento, Italy
[b] DCP, University IUAV of Venice, Italy

Keywords:
Children
Chances
Natural frequencies
Frequency format
Probabilistic reasoning
Bayesian reasoning

A B S T R A C T

Zhu and Gigerenzer (2006) showed that an appreciable number of Chinese children aged between 9 and 12 years old made correct quantitative Bayesian inferences requiring the integration of priors and likelihoods as long as they were presented with numerical information phrased in terms of natural frequencies. In this study, we sought to replicate this finding and extend the investigation of children's Bayesian reasoning to a different numerical format (chances) and other probability questions (distributive and rel- ative). In Experiment 1, a sample of Italian children was presented with the natural frequency version of five Bayesian inference problems employed by Zhu and Gigerenzer (2006), but only a tiny minority of them were able to produce correct responses. In Experiment 2, we found that the children's accuracy,as well as the coherence between their probability judgments, depended on the type of question but not on the format (natural frequency vs. chance) in which information was presented. We conclude that children's competence in drawing quantitative Bayesian inferences is lower than suggested by Zhu and Gigerenzer (2006) and, similarly to what happens with adults, it relies more on a problem representation that fosters an extensional evaluation of possibilities than on a specific numerical format.

1. Introduction

What kinds of probabilistic inferences emerge at earlier ages and what others require, instead, more time to be mastered? Whenare children able to engage in proper Bayesian reasoning? Does this depend on the way in which the relevant information is presented?

Converging empirical evidence indicates that preverbal infants (Téglás, Girotto, Gonzalez, & Bonatti, 2007; Téglás et al., 2011) as well as nonhuman primates (Rakoczy et al., 2014) possess at least some basic probabilistic intuitions that enable them to make *implicit probability inferences* like those necessary to determine which of two elementary events is more likely to occur. For example, looking time suggests that 12-month-old infants correctly expect a yellow ball, rather than a blue one, to exit from a container in which three yellow balls and only one blue ball are bouncing (Tégláset al., 2011). Probabilistic competence has also been claimed to guide optimal choices already in 10- to 14-month-old infants (Denison & Xu, 2010, 2014), even if this result has been replicated only with children older than 4 years (Girotto, Fontanari, Gonzalez, Vallortigara, & Blaye, 2016).

Explicit qualitative probabilistic inferences like those required to make predictions in accordance with prior probability have been documented at a later stage of development (Brainerd, 1981; Yost, Siegel, & Andrews, 1962; Téglás et al., 2007/Study 3; see also Sobel, Tenenbaum, & Gopnik, 2004 and Griffiths, Sobel, Tenenbaum, & Gopnik, 2011 for similar investigations concerning children's capacity to make probabilistic inferences about the causal properties of objects). For example, Girotto and Gonzalez (2008) showed that 5-year-old children correctly judged that a black token was more likely than a white one to be drawn from an opaque bag containing four black round tokens, one black square token, and three white square tokens. At the same age, children are also able to update their judgments based on new evidence. In the above example, if the children were told the shape of the token (before the outcome of the extraction was revealed), they proved able to integrate this new piece of information into their judgment: that is, when the extracted token was round, they kept betting on black, while when it was square, they changed their bet to white. Similar results have been obtained with preliterate and prenumerate indigenous Maya groups living in remote areas of Guatemala (Fontanari, Gonzalez, Vallortigara, & Girotto, 2014). The findings that 5-year-old children as well as preliterate and prenumerate individuals can make sound

* Corresponding author at: CIMeC, University of Trento, Corso Bettini, 31, 38068 Rovereto (TN), Italy.
E-mail address: stefania.pighin@unitn.it (S. Pighin).

qualitative Bayesian inferences based on priors and/or likelihoods strongly suggest that such inferences do not depend on formal education. This is also coherent with the findings that abstract knowledge of number and basic numerical skills (i.e., comparing and adding numerical quantities) precede (Barth, La Mont, Lipton, & Spelke, 2005; Barth et al., 2006) and are independent from (Butterworth, Reeve, Reynolds, & Lloyd, 2008; Pica, Lemer, Izard,& Dehaene, 2004) schooling.

Before the acquisition of symbolic number knowledge, young children therefore seem able to make (implicit or explicit) qualitative probabilistic inferences by considering and comparing the various ways in which an outcome may occur (*extensional reasoning*, Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999). Such judgments express a more-less relationship between two quantities, and may rely on a bare enumeration of relevant possibilities without requiring precise calculations. *Quantitative inferences* are more complex than qualitative ones because they involve the ability to manipulate and estimate exact relative amounts (*proportional reasoning*). Accordingly, much of the research that has been conducted on children's capacity to calculate probabilities has been confined almost exclusively to that ability (for a review, see Bryant & Nunes, 2012). Despite minor disagreements, problems that require proportional reasoning prove to be hard for children up to the age of roughly 10 years, both when numerical informa- tion concerns odds (e.g. 1:2) and, to a greater extent, conventional fractions (e.g., 1/3; see Fischbein & Gazit, 1984; Fujimura, 2001; Noelting, 1980; Nunes & Bryant, 1996; Pitkethly & Hunting, 1996; Schwartz & Moore, 1998).

The present study focuses on the developmental course of children's performance in a more sophisticated reasoning ability. In particular, we are interested in understanding when children become able to draw proper *quantitative Bayesian* inferences which involve a numerical integration between different values, as priors and likelihoods. As far as we can tell from the literature, Zhu and Gigerenzer (2006) is the only study which has investigated this issue. A sample of children attending an ordinary elementary school in Beijing (China) was presented with a number of Bayesian problems whose content was suited to children. There follows an example (p. 289):

> Pingping goes to a small village to ask for directions. In this village, 10 out of every 100 people will lie. Of the 10 people who lie, 8 have a red nose. Of the remaining 90 people who don'tlie, 9 also have a red nose. Imagine that Pingping meets a group of people in the village with red noses. How many of these peo- ple will lie? out of_.

When problems were phrased in terms of frequencies (like the one above), the rate of Bayesian responses (averaged across two experiments and weighted by their sample sizes) provided by fourth, fifth, and sixth graders (aged between 9 and 12) were 18.7%, 39%, and 53.5%, respectively (p. 294). However, no child could provide any Bayesian response when problems were phrased in terms of percentages. Zhu and Gigerenzer (2006) interpreted their results as supporting the hypothesis that the human mind is not designed for probabilities or percentages, but needs natural frequencies[1] to make sound Bayesian inferences (Gigerenzer, 1996; Gigerenzer & Hoffrage, 1995, 1999).

Zhu and Gigerenzer's (2006) study has had a considerable impact on the literature. The result that a substantial number of children aged between 9 and 12 years old make sound Bayesian inferences when provided with natural frequency but not single-event probability information has been interpreted as crucial evidence supporting the claim that natural frequencies enable humans to reason the Bayesian way, while other formats prevent it (Brase, 2008; Galesic, Gigerenzer, & Straubinger, 2009; Gigerenzer, 2015; Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz,& Woloshin, 2008). However, to our knowledge, these results have never been replicated.[2]

Given the importance of Zhu and Gigerenzer's (2006) results not only from a developmental perspective but also for their more general implications in regard to reasoning research, confirmatory experimental evidence is conspicuous by its absence. All the more so because the success rate of Zhu and Gigerenzer's (2006) Chinese sixth (and, to a certain extent, fifth) graders exceeds that typically found with naïve and even most educated Western adults (Bramwell, West, & Salmon, 2006; Girotto & Gonzalez, 2001; Hoffrage & Gigerenzer, 1998). For example, in Bayesian problems phrased in terms of natural frequency, accuracy rates of 46% have been reported with Austrian university students (Gigerenzer & Hoffrage, 1995) and experienced German physicians (Hoffrage & Gigerenzer, 1998), 31% with US undergraduates (Sloman, Over, Slovak, & Stibel, 2003, Exp. 1b), below 25% with patients at Spanish hospitals (Garcia-Retamero & Hoffrage, 2013), 2% with US adults recruited using Amazon Mechanical Turk (Pighin, Gonzalez, Savadori, & Girotto, 2016), and 0% with UK midwifes recruited at training events or in maternity wards (Bramwell et al., 2006). Our Experiment 1 was therefore aimed at replicating Zhu and Gigerenzer's (2006) study concerning problems phrased in a natural frequency format, using a different sample of children of the same age (i.e., fourth, fifth, and sixth graders) from another country (Italy).

We also wanted to empirically assess Zhu and Gigerenzer's (2006) conclusion about the advantage of natural frequencies over single-event probabilities on children's Bayesian reasoning. The hypothesis that frequencies could facilitate probabilistic reasoning has a long (and not always linear) history. It originated from the observation that the rate of some well-known biases (e.g., the conjunction fallacy) reduced when the problems were framed in terms of frequencies (Cosmides & Tooby, 1996; Gigerenzer, 1991; Gigerenzer & Hoffrage, 1995, 2007; Gigerenzer, Todd, & ABC Research Group, 1999, but a preliminary version of this hypothesis was already put forward in Tversky & Kahneman, 1983, p. 309). Such a hypothesis, however, has been increasingly challenged by a growing body of evidence showing that frequencies are not inherently easier to process than percentages (Cuite, Weinstein, Emmons, & Colditz, 2008; Waters, Weinstein, Colditz, & Emmons, 2006) and that, once various confounding factors have been eliminated, their advantage disappears as well (Evans, Handley, Perham, Over, & Thompson, 2000; Reyna & Brainerd, 2008; Sloman et al., 2003; Tentori, Bonini, & Osherson, 2004). The advo- cates of the frequentist hypothesis have rejected these results because they were obtained with problems in which frequencies were typically normalized, and they have reiterated their position with respect to Bayesian updating problems (like, for example, the

[1] Note that, as pointed out by an anonymous reviewer of this paper, the term *natural* does not reflect any biological claim, but only a speculation about an alleged evolutionary advantage. According to this, natural frequencies constitute a cognitively privileged format because they represent the outcomes of the process of counting and classifying the occurrences of events as they are experienced (*natural sampling*, Kleiter, 1994). On the other hand, the human mind ''would not be tuned to probabilities or percentages as input format'' (Gigerenzer & Hoffrage, 1995, p. 686; but see also Gigerenzer, 1996), because these do not correspond to the typical way in which humans have dealt with statistical information over their evolution.

[2] Luecking (2004) and Multmeier (2012) have been repeatedly mentioned (e.g., by Gigerenzer, 2008; Gigerenzer, 2015; Martignon & Kuntze, 2015) as supporting Zhu and Gigerenzer's (2006) conclusions. However, both these studies are unpublished. Multmeier (2012) is available online, and we note that the data reported therein are only partially coherent with those of Zhu and Gigerenzer (2006). Indeed, Multmeier's first experiment, the only one whose stimuli and participants are directly comparable with those of Zhu and Gigerenzer, did not involve fifth or sixth graders, and the accuracy rate of fourth graders seemed somewhat lower (13%) than that of Zhu and Gigerenzer's (2006) participants.

Pingping problem above), which are suitable to be framed in terms of natural frequencies. Whereas the advantage of natural frequencies over percentages in these kinds of problems seems less controversial, the same is not true for its interpretation. Indeed, when framed in terms of natural frequencies, such problems become undeniably simpler, since the correct posterior probability can be provided without applying the Bayes rule. According to the *nested set* view (e.g., Barbey & Sloman, 2007; Evans et al., 2000; Fox & Levav, 2004; Girotto & Gonzalez, 2001; Johnson-Laird et al., 1999; Sloman et al., 2003), the real source of facilitation is not the frequency format but the partition of the data into exhaustive subsets that explicitly provides the responders with the relevant conjunctive events (referred to as *partitive* or *partitioned* structure, Girotto & Gonzalez, 2001; Macchi, 1995). In similar vein, according to the *Fuzzy Trace Theory* (Brainerd & Reyna, 1990; Reyna, 1991; Reyna & Brainerd, 2008), natural frequencies (as well as any manipulation that explicitly represents parts and wholes, such as Venn diagrams or two-by-two tables) simplify processing because they allow relevant classes to be represented discretely and, possibly, combined in various probability judgments (e.g., conjunctions or conditional probabilities). When, instead, classes are overlapping (as in the percentages formulation of Bayesian updating problems), reasoners would provide erroneous judgments by focusing on numerators to the detriment of part-whole relations. The frequentist view's defenders (e.g., Gigerenzer & Hoffrage, 1995) objected that such a nested-set or class-inclusion structure is intrinsically linked to the natural frequency format and cannotbe implemented with single-event probabilities. In response, Girotto and Gonzalez (2001) reproduced the same structure (and consequent facilitation) by means of *numbers of chances*[3] (already appeared in Johnson-Laird et al., 1999). As occurs with natural frequencies, numbers of chances can state probabilities as positive integer numbers, can be arranged in subset relations, and can preserve the sample size of the reference class, and therefore can be used ina partitioned representation of a Bayesian problem. The nature of numbers of chances has long been debated (see, for example, Brase, 2008; Girotto & Gonzalez, 2002; Hoffrage, Gigerenzer, Krauss, & Martignon, 2002). In a recent study, we demonstrated that laypeople interpret them as single-event probabilities, with no detri- ment to reasoning accuracy (Pighin, Tentori, & Girotto, in press). Since Zhu and Gigerenzer themselves (2006, p. 303) acknowledged that their results did not exclude the possibility that other numerical representations could foster insights leading to the same computa- tional facilitation of natural frequencies, in Experiment 2, we thought it would be of interest to explore the impact of number of chances on children's Bayesian inferences. Moreover, we introduced other probability questions. As highlighted by Girotto and Gonzalez (2001), the specific question presented to participants affects the accuracy of their probability judgments, independently from the information format. In particular, in addition to the *standard* proba- bility question used by Zhu and Gigerenzer (2006), we posed our participants also a *distributive* question (i.e., to apportion the number of observations or chances in favor of and against a given hypothe- sis). This kind of question does not require normalization and is known to increase the accuracy rate of quantitative Bayesian infer- ences in adults (Girotto & Gonzalez, 2001; Pighin, Gonzalez, Savadori, & Girotto, 2015). Finally, we checked whether children understand a basic implication of their (standard and/or distributive) probability judgments. In fact, probability judgments are generally assumed to reveal individuals' expectations about the truth or falsity of the hypotheses at issue. This means that when a person assesses $pr(H|E)$ > [<] 0.5 or $pr(H\&E)$ > [<] $pr(\ H\&E)$, s/he is expected to believe H to be more [less] likely than H in light of E. To the bestof our knowledge, this (very basic) assumption has never been tested, neither with adults nor with children. For this reason, we included in our stimuli a final *relative* question concerning the direct comparison between $pr(H|E)$ and $pr(:H|E)$.

2. Experiment 1

In Experiment 1, we investigated whether the accuracy of a sample of Italian children in solving Bayesian problems phrased in terms of natural frequencies was comparable with the accuracy of Zhu and Gigerenzer's (2006) Chinese participants of the same age.

2.1. Method

2.1.1. Participants

Fourth (n = 34; M_{age} = 9 years and one month, SD = 0.34; 18 females), fifth (n = 43; M_{age} = 10 years and one month, SD = 0.32; 23 females), and sixth (n = 43; M_{age} = 11, SD = 0.41 years and two months; 21 females) graders from an ordinary elementary school in a north-east Italian region participated in the experiment. Allof them were native Italian speakers. Signed consent was obtained from their parents.

2.1.2. Materials and design

Zhu and Gigerenzer (2006) administered various problems to their participants (more precisely, seven problems in Experiment1 and ten in Experiment 2). To limit attentional and fatigue effects, we preferred to give our participants a smaller number of prob- lems. Participants' success rate for each single problem was not reported in Zhu and Gigerenzer (2006). To select the subset of problems to be employed in Experiment 1, we then ran a pilot study involving 40 US adults (M_{age} = 30 years, SD = 5.6; age range 22–44 years; 14 women), who were recruited using the Amazon Mechanical Turk (AMT) platform. Participants received a payment of $2 to solve the ten problems used by Zhu and Gigerenzer (2006; problems reported in Appendix A). The presentation order of the problems was randomized. Responses were classified using the *strict outcome criterion* also employed by Zhu and Gigerenzer (2006), according to which a response is considered correct only if it exactly coincides with the Bayesian solution. The percentages of correct answers in the ten problems are reported in Appendix A: they ranged between 17.5% (in Problem 4) and 37.5% (in Problem1), with an average of 27%. Note that, despite identical inference problems, such accuracy rates yet again appear lower than those of Zhu and Gigerenzer's (2006) fifth and sixth grade participants.

For our Experiment 1, we selected the five problems that received the greatest number of correct responses in the pilot (i.e., problems 1, 7, 8, 9, and 10 in Appendix A). Children of all ages were given 30 min to solve the five problems, whose presentation order was counterbalanced across participants.

To avoid possible anxiety caused by the presence of an unknown person (i.e., the experimenter), the children were tested by their teachers, who were instructed not give them any help. The experiment was administered in classrooms and introduced as part of the regular teaching activity. The children were informed that the questionnaires were anonymous and that their performances would not be evaluated. The problems were preceded by the following sentence:

> "Please solve the following problems. Each of them includes some numbers and a question."

[3] According to Girotto and Gonzalez (2001, 2002), probabilities can be expressed in terms of chances. For instance, we may say that in a regular deck, there are 4 chances out of 52 of drawing a Queen. Such number of chances expresses theoretical possibilities that incorporate abstract knowledge (such as the information that the deck is a regular one, in the aforementioned example), and does not necessarily result from a sequential process of observing and counting past events.

2.2. Results

All participants completed the task well within the 30-min time limit. As in both Zhu and Gigerenzer's (2006) study and our pilot, only responses that exactly matched the Bayesian solution of the problems were considered correct. Table 1 shows the number of participants who gave (from 0 to 5) Bayesian responses in each age group. These were not uniformly distributed: independently from the age group, the large majority of children (75%) solved none of the problems. Sporadic Bayesian responses were observed in each age group, but only a minority of sixth graders (19%, which corresponds to 7% of all participants) solved all the problems. The overall rates of Bayesian responses (out of 5 X the number of participants in the specific group) significantly increased with age(T_{JT} = 2.09, p = 0.036) and were 7%, 11%, and 28% for fourth, fifth, and sixth graders, respectively.

The overall rate of Bayesian responses in the three age groups weighted by their sample sizes was 16%. If we compare this value with those reported in Zhu and Gigerenzer's (2006) study (38% and 36% in Experiment 1 and 2, respectively) they appear to be largely heterogeneous ($Q(2) = 94.92$, $p < 0.001$) and inconsistent ($I^2 = 98\%$; see Higgins, Thompson, Deeks, & Altman, 2003). High levels of inconsistency are also obtained on comparing the rates of Bayesian responses within each age group (see Table 2 for detailed results).

We also analyzed children's responses that did not match the correct solution of the problems by classifying them into one of the following categories, which summarize various non-Bayesian strategies observed with adults (e.g., Gigerenzer & Hoffrage, 1995), some of which (see points a, b, c, and d below) have been also employed by Zhu and Gigerenzer (2006):

(a) ''Base-rate only,'' in which only the prior probability is reported, while the new evidence is disregarded (i.e., $pr(H)$, for example ''10 out of 100'' in Problem 1);
(b) ''Evidence-only,'' which focuses on the occurrence of the evidence among all cases (i.e., $pr(E)$, for example ''17 out of 100'' in Problem 1);
(c) ''Representative thinking,'' which reports how often the evidence occurs when the hypothesis is true (i.e., $pr(E|H)$, for example ''8 out of 10'' in Problem 1);
(d) ''Pre-Bayes,'' which reports the right denominator but associated with a wrong numerator (e.g., ''10 out of 17'' or ''2 out of 17'' in Problem 1);
(e) ''Joint occurrence,'' which reports how often both the evidence and the hypothesis occur among all cases (i.e., $pr(H\&E)$, for example, ''8 out of 100'' in Problem 1; note that the answers driven by this strategy are correct as regards the numerator but not the denominator).

Answers that escaped the above categories were classified as ''Other'' (e.g., responses like ''8 out of 9'', ''27 out of 100'', or ''73 out of 100'' in Problem 1).

Table 3 reports the distribution of the children's non-Bayesian responses. The majority of them fell into the ''Other'' category (55% across different age groups), suggesting that, differently fromadults' errors, the children's incorrect responses were mainly ran- dom (e.g., summing or subtracting values arbitrarily picked from the problem text). A similar pattern of results was obtained alsoin Zhu and Gigerenzer (2006), with 63% of non-Bayesian responses, which fell outside the four categories that they considered (i.e., a, b, c, and d in the list above). However, the strategy of relying only on the prior probability (''Base-rate only'' strategy) proved to be the most systematic error for all age groups of our participants ($\chi^2(4)$ = 173.7, $p < 0.001$), while the most common non-Bayesian strategy among Zhu and Gigerenzer's (2006) participants was the ''Pre-Bayes'' one ($\chi^2(3)$ = 104.9, $p < 0.001$).

If compared to typical adults' non-Bayesian responses, the ''Base-rate only'' strategy contrasts with the ''Base-rate neglect'' commonly observed when numerical information is presented using percentages (Barbey & Sloman, 2007; Gigerenzer & Hoffrage, 1995), but it is perfectly in line with the results of variousstudies concerning Bayesian inference problems framed in a fre- quentist format (Evans et al., 2000; Hoffrage & Gigerenzer, 1998; Pighin et al., 2016). This finding has been interpreted by Evanset al. (2000, pp. 211–212) as a difficulty in calculating the right denominator, which is not made explicit in the information pro- vided to participants. Such interpretation, however, does not account for ''Base-rate only'' responses which involve also a wrongnumerator (as happens when false negatives are possible). Anothertentative explanation is that (at least some) participants conflate $pr(H|E)$ with $pr(H\&E)$. Indeed, in the absence of possible false nega- tives, $pr(H\&E)$ equals to $pr(H)$, and therefore the confusion between $pr(H|E)$ and $pr(H\&E)$ would produce answers which are numerically identical to the base-rate. In agreement with this, the ''Base-rate only'' strategy is the adults' prevalent error when false negatives are not possible (as in Evans et al., 2000; Pighin et al., 2016);

Table 1
Number of participants (n) for each of the three age groups and for all of them (Overall) who gave zero, one, . . ., five Bayesian responses in Experiment 1. Among fourth graders, for example, 28 children did not solve any problem, three children solved one problem, one child solved two problems, and so on. The total (and corresponding percentages) of Bayesian responses for each age group are also indicated.

Grade	N	Number of Bayesian responses						Total
		0/5	1/5	2/5	3/5	4/5	5/5	
Fourth	34	28	3	1	1	1	0	12 (7%)
Fifth	43	33	4	2	1	3	0	23 (11%)
Sixth	43	29	0	2	0	4	8	60 (28%)
Overall	120	90	7	5	2	8	8	95 (16%)

Table 2
Proportions of Bayesian responses (and corresponding percentages) for each age group and for all participants (Overall), in Zhu and Gigerenzer's (2006) Experiments 1 and 2 (Z&G 1 and Z&G 2, respectively) and in our Experiment 1. The last two columns report heterogeneity (Cochran's Q) across the three experiments and the percentage of variation across experiments due to heterogeneity rather than chance (I^2).

Grade	Z&G 1	Z&G 2	Experiment 1	Cochran's Q (df = 2)	I^2
Fourth graders	19/112 (17%)	58/300 (19%)	12/170 (7%)	15.07, $p < 0.01$	87
Fifth graders	31/105 (29%)	127/300 (42%)	23/215 (11%)	69.87, $p < 0.001$	97
Sixth graders	69/98 (70%)	144/300 (48%)	60/215 (28%)	54.43, $p < 0.001$	96
Overall	119/315 (38%)	329/900 (37%)	95/600 (16%)	94.92, $p < 0.001$	98

Table 3
Number of non-Bayesian responses (and corresponding percentages) falling into the four categories employed by Zhu and Gigerenzer (2006), ''Joint occurrence", unspecified types (''Other"), and missing responses (''Missing"), for each of the three age groups and for all participants (''Overall") of Experiment 1. The last row shows the numbers of non- Bayesian responses (and corresponding percentages) in Zhu and Gigerenzer's (2006) Experiments 1 and 2 (Z&G). Note that Zhu and Gigerenzer (2006) provided only aggregate numbers for ''Joint occurrence", ''Other" and ''Missing" responses.

Grade	Non-Bayesian responses						Missing
	Base-rate only	Evidence-only	Representative thinking	Pre-Bayes	Joint occurrence	Other	
Fourth	53 (33%)	8 (5%)	14 (9%)	2 (1%)	9 (6%)	66 (42%)	6 (4%)
Fifth	29 (15%)	15 (8%)	13 (7%)	2 (1%)	5 (3%)	119 (62%)	9 (4%)
Sixth	34 (22%)	10 (6%)	6 (4%)	1 (1%)	10 (6%)	92 (59%)	2 (2%)
Overall	116 (23%)	33 (7%)	33 (7%)	5 (1%)	24 (5%)	277 (55%)	17 (3%)
Z&G	65 (8%)	56 (7%)	22 (3%)	140 (18%)		484 (63%)	

otherwise, non-Bayesian responses seem to be equally distributed between $pr(H)$ and $pr(H\&E)$ (Pighin, Tentori, Savadori, & Girotto, 2017). As said above, our children participants' non-Bayesian responses mainly targeted $pr(H)$, even if false negatives were possible. This may have been due, in part, to processing difficulties (see on this, Reyna & Brainerd, 1995a, 1995b; Reyna, 2004) that prompt children to take the first numerical value presented inthe problem (which is exactly $pr(H)$). Further research could explore this hypothesis more in depth. What it is relevant for the purposes of this study is that, regardless of the specific errors and their underlying root causes, we failed to replicate Zhu and Gigerenzer's (2006) results concerning the accuracy rate of fourth,fifth and sixth graders in Bayesian inference problems. Coherently with what has been observed with Westerns adults, only a tiny number of Italian children in the three age groups considered wereable to solve the problems, even if they were phrased in a natural frequency format.

3. Experiment 2

Experiment 2 was in the first place aimed at extending previous investigations on children's capacity to solve Bayesian inference problems by comparing the effects of different information formats (natural frequencies vs. chances), that keep the subset relations equally transparent.

Secondly, we wanted to explore the accuracy of probability judgments elicited by different probability questions, and possible relations among them. As highlighted by Girotto and Gonzalez (2001), the effect of information format (and corresponding problem representation) may be mediated by the type of question presented to participants (see also Barbey & Sloman, 2007). Thus, along with the standard question, which asked participants to assess the probability of H in light of E, in Experiment 2, we presented participants also with a distributive question,[4] which asked them to apportion the number of cases in favor of H and $\neg H$ (Girotto& Gonzalez, 2001). The standard and the distributive judgments dif-fer in their computational complexity, since the former but not the latter implies the ability to see the same subset (i.e., $H\&E$) simultaneously as a separate part and as an element contributing to the whole (i.e., the disjunction of $H\&E$ and $\neg H\&E$). Put otherwise, a standard judgment concerns a *part-whole relationship* that requires normalizing $H\&E$ by E, while a distributive question involves a *part-part relationship* that can be handled by comparing $H\&E$ and $\neg H\&E$ (for more on the difficulty of these kinds of comparisons, see Reyna & Brainerd, 1995a, 1995b; Spinillo & Bryant, 1991). Distributive judgments can be considered Bayesian inferences that rely on extensional reason- ing. We therefore expected to obtain a greater number of correct distributive judgments than standard ones. Moreover, in order to test the (often tacit) assumption that participants' relative judgment can be inferred from their standard or distributive judgment, we introduced a final explicit relative question, which asked partici- pants to select which between H and $\neg H$ was more likely in light of E. Relative judgments are explicit qualitative probabilistic judg- ments which express the comparison between two conditionalprobabilities.

Finally, to facilitate the children's reasoning, we also made some changes in the problems (see Table 4 for an illustrative summary). Like the majority of Bayesian inference problems, Zhu and Gigerenzer's (2006) ones displayed numerical values which expressed $pr(H)$, $pr(E\&H)$, and $pr(E\&\neg H)$. It is well known that, in dealing with formal problems, children tend to use all available numbers to arrive at a solution even if these are not necessary, with a consequent possible increase in errors (Low & Over, 1993; Marzocchi, Lucangeli, De Meo, Fini, & Comoldi, 2002). Since both natural frequencies and number of chances already incorporate the base rate, we omitted $pr(H)$. We also included the information concerning the probability of the evidence (i.e., $pr(E)$), which otherwise would need to be calculated in order to answer the standard question. Moreover, we explicitly provided in the text the two relevant values $pr(H|E)$, and $pr(\neg H|E)$, in place of the corresponding likelihoods. A different simplified Bayesian inference problem (named ''short information menu") was already proposed by Gigerenzer and Hoffrage (1995), with no significant improvement in performance. In that case, the numerical information displayed was $E\&H$ and $E\&\neg H$ (or, alternatively, $E\&H$ and E). While such changes could simplify the calculus of the standard judgment, ours make it possible to identify the correct standard judgment without any calculation at all.

3.1. Method

3.1.1. Participants

As in Experiment 1, participants were children from an ordinary elementary school in a north-east Italian region. Again, fourth

Table 4
Illustrative summary of the changes introduced in the problems employed in Experiment 2. The first column reports the numerical information typically presented in Bayesian inference problems, while the second column reports the numerical information presented in our simplified version.

Traditional version		Simplified version	
$pr(H)$	You met 100 persons in this village, 10 of them were liars	$pr(E)$	You met 100 persons in this village, 24 of them had a red nose
$pr(E\&H)$	8 of the 10 persons who were liars had a red nose	$pr(H\|E)$	15 of those 24 persons were liars
$pr(E\&\neg H)$	9 of the remaining 90 persons who were not liars also had a red nose	$pr(\neg H\|E)$	The remaining 9 persons were not liars

[4] Note that, when problems are phrased in natural frequencies or numbers of chances, standard and distributive judgments convey the same probabilistic information, and indeed one can be easily inferred from the other.

($n = 43$; $M_{age} = 9$ years and eight months, $SD = 0.26$; 19 females), fifth ($n = 51$; $M_{age} = 10$ years and nine months, $SD = 0.32$; 22 females), and sixth ($n = 65$; $M_{age} = 11$ years and eight months, $SD = 0.29$; 38 females) graders were involved.

3.1.2. Materials and design

Procedure and instructions were the same as those of Experiment 1. Each participant was asked to solve four Bayesian inference problems (see Appendix B). Problems 1 and 2 resembledthose used by Zhu and Gigerenzer (2006), while problems 3 and 4 resembled the disease problems used in Girotto and Gonzalez (2001) and Pighin et al. (2015), with content adapted for children. The order of the four problems was counterbalanced across participants.

For each problem, participants were asked to answer a distributive, a standard, and a relative question, always in this order (see Appendix B). The order of the questions was kept constant across participants as we preferred them to face first the distributive judgment (which we expected to be easier), so that they could use it to tackle the following (more difficult) standard judgment. The relative question was posed at the end because we were inter-ested in understanding whether the children's distributive and standard judgments translated into coherent relative judgments more than in the overall accuracy rate of the latter.

The children were randomly assigned to either a natural frequency or a chance version. In the natural frequency version, the information was conveyed in terms of past observations, while it focused on a single current event in the chance version. As said above, differently from Experiment 1, the numerical values for $pr(E)$, $pr(H|E)$, and $pr(H|\neg E)$ were explicitly provided, so that all the questions could be answered without any computation. This was expected to strongly decrease the computational demand of the problems, with a consequent increase in the accuracy rate.

3.2. Results

All participants completed the four problems well within the 30-min time limit. Participants' accuracy was analyzed according to the same criterion employed in Experiment 1. The accuracy rates for the four problems were compared by three Friedman tests (one for each distributive, standard, and relative question) and resulted not statistically different (all $p > 0.05$). Accordingly, we aggregated participants' responses across problems. Table 5 reports the number of participants who gave (from 0 to 4) Bayesian responses for the three questions in each age group of Experiment 2.

Accuracy rates tended to increase with age only in the distributive ($T_{JT} = 2.85$, $p < 0.01$), and in the relative ($T_{JT} = 2.67$, $p < 0.01$) questions, but not in the standard one ($T_{JT} = 1.64$, $p = 0.10$). More importantly, as in Experiment 1, when information was phrased in terms of natural frequencies, a conspicuous number of children could not provide any correct standard judgment (57%, 37%, and 43%, for fourth, fifth, and sixth graders, respectively), and only a minority of them correctly solved all the problems (9%, 30%, and 23%, for fourth, fifth, and sixth graders, respectively).

As predicted, participants' accuracy rates were very similar for the two information (natural frequency vs. chance) formats. Fifth and sixth graders' Bayesian natural frequency and chance responses did not differ for any of the questions (all $p > 0.05$, see Table 3 for details). Among fourth graders, the only significant difference was obtained for relative judgments, which were more accurate in the chance than natural frequency version ($Mdn = 3$ and $Mdn = 2$, respectively; $U = 107$, $p = 0.002$). To further investi- gate the effect of the information type on the children's Bayesian inferences, an aggregate analysis was performed by pooling partic- ipants' responses across ages. The results showed that the number of correct distributive and relative judgments was greater in the chance than natural frequency version ($U = 2461$, $p = 0.011$ and

Table 5
Number of participants (n) for each of the three age groups and for each of the three questions who gave zero, one, . . ., four Bayesian responses in Experiment 2. The last two columns report the median number of correct responses (Mdn) and the p-values for Mann-Whitney U test comparisons between the natural frequency and chance versions.

Grade	n	Number of Bayesian responses					Total	Mdn	p-value
		0/4	1/4	2/4	3/4	4/4			
Distributive question									
Fourth	43	16	3	9	5	10	76 (44%)	2	
Natural frequency	21	10	3	3	2	3	27 (32%)	1	0.063
Chance	22	6	0	6	3	7	49 (56%)	2	
Fifth	51	15	2	1	10	23	126 (62%)	3	
Natural frequency	27	11	1	0	5	10	56 (52%)	3	0.107
Chance	24	4	1	1	5	13	70 (73%)	4	
Sixth	65	13	4	4	11	33	177 (68%)	4	
Natural frequency	30	7	2	3	5	13	75 (63%)	3	0.258
Chance	35	6	2	1	6	20	102 (73%)	4	
Standard question									
Fourth	43	24	6	3	3	7	49 (28%)	0	
Natural frequency	21	12	3	2	2	2	21 (25%)	0	0.677
Chance	22	12	3	1	1	5	28 (32%)	0	
Fifth	51	21	6	2	5	17	93 (46%)	1	
Natural frequency	27	10	3	1	5	8	52 (48%)	2	0.780
Chance	24	11	3	1	0	9	41 (43%)	1	
Sixth	65	29	3	6	4	23	119 (46%)	2	
Natural frequency	30	13	3	5	2	7	47 (39%)	1	0.354
Chance	35	16	0	1	2	16	72 (51%)	3	
Relative question									
Fourth	43	1	6	10	12	14	118 (69%)	3	
Natural frequency	21	1	5	7	5	3	46 (55%)	2	0.002
Chance	22	0	1	3	7	11	72 (82%)	3	
Fifth	51	2	4	6	17	22	155 (76%)	3	
Natural frequency	27	2	4	3	6	12	76 (70%)	3	0.415
Chance	24	0	0	3	11	10	79 (82%)	3	
Sixth	65	1	3	11	12	38	213 (82%)	4	
Natural frequency	30	0	3	4	8	15	95 (79%)	3	0.290
Chance	35	1	0	7	4	23	118 (84%)	4	

$U = 2402.5$, $p = 0.005$, respectively). Instead, no significant effect of the information format was observed for standard judgments ($p = 0.557$).

Despite the fact that the values required to answer both questions were explicitly provided in the text, the distributive question proved simpler than the standard one. In fact, for the former, we obtained 44%, 62%, and 68% of correct responses among fourth, fifth and sixth graders, respectively, while for the latter the proportions were 28%, 46% and 46%, respectively. Separate Wilcoxon tests for each age group indicated that there was a statistically significant difference between the proportion of correct distributive and standard judgments ($Z = 2.60$, $p = 0.009$, $Z = 3.20$, $p = 0.001$, and $Z = 4.185$, $p < 0.001$ for fourth, fifth and sixth graders respectively).

The accuracy of relative judgments was rather good: 69%, 76%, and 82% for fourth, fifth, and sixth graders, respectively (all significantly higher than 50%, $p < 0.001$, binomial test). A logistic regression analysis indicated that the distributive and standard judgments as a set reliably distinguished between judgments which endorsed or did not endorse the more likely hypothesis ($V^2(2) = 94.28$, $p < 0.001$). However, under the Wald criterion, only the accuracy of the distributive judgments made a significant contribution to the prediction of the accuracy of relative judgments ($p < 0.001$). To further investigate whether the children understood the implications of their (standard and/or distributive) probability judgments, we considered the agreement ‘‘in direction” between distributive and relative judgments and between standard and relative ones. That is, independently from their accuracy, we counted the number of distributive and standard judgments in favor of H : [H] which were associated with corresponding relative judgments endorsing H [: H]. The agreement with relative judgments was generally good in both cases, but it was higher for distributive than standard judgments (83% vs. 72%, $V^2(1) = 21.7$, $p < 0.001$). Accordingly, relative judgments following incoherent distributive and standard judgments (i.e., a distributive judgment in favor of H and a standard judgment in favor of : H, or vice versa) were more likely to be in the same direction as distributive than standard judgments (70% vs. 30%, $p < 0.001$, binomial test). We also considered the sub-set of incoherent distributive and standard judgments in which one of the two was correct, and found that relative judgments were more likely to be correct when preceded by correct distributive than standard judgments (85% vs. 57%, $V^2(1) = 9.38$, $p = 0.002$).

To sum up, although the computational difficulty of the problems used in Experiment 2 was lower than that of the problems typically employed in the literature, the Italian children's standard probability judgments proved to be independent from the information type (i.e., natural frequency vs. chance) and, yet again, less accurate than suggested by Zhu and Gigerenzer's (2006) study. Distributive judgments were consistently more accurate than standard ones, as well as being more aligned with the children's relative judgments. Moreover, they appeared to depend (at least in part) on the information type, with a slight advantage for the chance format.

4. Discussion

Once symbolic number knowledge has been acquired, are children able to draw correct Bayesian inferences which require the integration of priors and likelihoods? The only published experimental study on this topic offers an optimistic answer. According to Zhu and Gigerenzer (2006), an appreciable number of children are able to solve Bayesian problems as long as the relevant information is provided to them in terms of natural frequencies. This finding has had a wide impact, although it has not been independently confirmed and is largely inconsistent with Western adults' modest performance in similar problems. The research reported in this paper was aimed at replicating and extending Zhu and Gigerenzer's (2006) study. In Experiment 1, a sample of Italian children was asked to solve the natural frequency version of five Bayesian inference problems employed by Zhu and Gigerenzer (2006). Although we selected the problems with the greatest accuracy rate in the pilot experiment with adults, only a tiny minority of the Italian children correctly solved them, so that our results failed to reproduce those of Zhu and Gigerenzer (2006). In Experiment 2, we broadened the investigation of the children's Bayesian reason- ing to include new simplified problems, whose relevant informa- tion was phrased in different numerical formats (natural frequency and chance), and which employed various (namely, dis- tributive, standard, and relative) probability questions. We found that the children's performance generally improved, without any significant advantage for natural frequencies over chances. Results of Experiment 1 and 2 converge, therefore, in disconfirming the hypothesis that natural frequencies are a cognitively privileged format (Gigerenzer, 1996; Gigerenzer & Hoffrage, 1995; Hoffrage & Gigerenzer, 2004), and indicate that children can reason about single-event probabilities when these are presented so as to foster an extensional evaluation of possibilities (Johnson-Laird et al., 1999). One could wonder if such facilitation depends on numbers of chances to mimic the structure of natural frequencies (e.g., (Gigerenzer, 2003). However, recent evidence suggests that chances are perceived by laypeople as being distinct from natural frequencies, and that they have a facilitatory effect on Bayesian inference tasks that is completely independent from their (minor) frequentist readings (Brase, Pighin, & Tentori, in press; Pighin et al., in press).

In all age groups, distributive judgments were more accurate than standard ones, although, in the simplified Bayesian problems employed in Experiment 2, the relevant values for forming both judgments were explicitly provided in the text (i.e., there was no need to make any calculation). The advantage of distributive judgments over standard ones is not surprising. As said, tasks that require consideration of part-whole relationships are more demanding than those that require consideration of part-part relationships (Reyna & Brainerd, 1995a, 1995b; Spinillo & Bryant, 1991), and children understand proportions as odds (e.g., 1:2) before they understand them as fractions (e.g., 1/3; Nunes & Bryant, 1996). Finally, the majority of relative judgments werecoherent (in direction) with the participants' previous judgments, providing general support for the assumption that standard and distributive judgments (regardless of their accuracy) reveal indi- viduals' expectations about the truth or falsity of the hypotheses under consideration. However, relative judgments were better pre-dicted by distributive judgments than standard ones. As already claimed by Acredolo, O'Connor, Banks, and Horobin (1989), ‘‘chil- dren do have the ability to choose on the basis of odds . . . [while] they may lack the ability to generate precise mathematical solu- tions” (pp. 943–4). When problems are phrased in natural frequen-cies or numbers of chances, the qualitative information about the relative probability of the two hypotheses at issue is made explicit by the distributive judgment, while it can be derived from the stan-dard judgment only by computing the ratio between the numera- tor and the denominator and comparing it with 0.5. Therefore, a correct distributive judgment can be a reliable basis for a correct relative judgment even when the corresponding quantitative standard judgment is inaccurate. The greater agreement between relative and distributive (rather than standard) judgments is also compatible with the large body of evidence showing that people often fail to take adequate consideration of alternative hypotheses (Beyth-Marom & Fischhoff, 1983; Hayes et al., 2015; Skov & Sherman, 1986). Indeed, traditional standard questions may boost focal biases because (unlike distributive or relative ones) they concentrate respondents' attention on a single hypothesis, to the detriment of the specific information concerning alternative hypotheses.

Explaining the source of the gap between our Italian participants and Chinese age-mates tested by Zhu and Gigerenzer (2006) falls outside the scope of this study. It is well known that different language systems, learning environments, and cultural varieties can shape the acquisition and the representation of numerical concepts (Ansari, 2008; Campbell & Xue, 2001; Posner & Rothbart, 2005; Torbeyns, Schneider, Xin, & Siegler, 2015). It has also been suggested that Chinese and Western speakers may differ in the brain representations of number processing (Tanget al., 2006). Although the timing in which students are introduced to fraction and ratio concepts is similar in China and Italy (and in line with that of other Western countries, see also Torbeyns et al., 2015), Chinese students' higher achievement in mathematics is not a matter of dispute. Their performance typically exceeds the international average (with which Italian students are aligned) by the equivalent of nearly three years of schooling (OECD, 2015). What linguistic factors and/or numerical competences could bring about the observed differences in Bayesian reasoning, as well as their possible relative weight, remains to be determined. In any case, if confirmed, this gap would strongly limit the generalizability of the empirical results concerning reasoning experiments from Western to Eastern participants (especially but not only in developing ages), and vice versa. Of course, other factors could have generated or at least contributed to the different results obtained in the two studies. Unfortunately, Zhu and Gigerenzer (2006) did not specify some aspects concerning their data collection that could be extremely relevant, such as, for example, whether ques- tioners were anonymous and/or whether the children expectedto be evaluated for their performances. Therefore, the effects of social pressure factors, especially the need for achievement and competitiveness, could not be ruled out as well.

Uncertainty pervades the most important aspects of people's lives, affecting their social relations, health judgments, economic decisions, and many other important forms of behavior. Studying how probabilistic reasoning evolves is highly informative on how people develop the basic intuitions that enable them to deal with randomness, to consider possible outcomes of alternative courses of action, to recognize the connections between events, and many other cognitive skills that provide the foundation for learning and intelligence. The traditional Piagetian view that young children completely lack the logical capacity to take probabilities into account (Piaget & Inhelder, 1975) has been disproved by a number of empirical results (see, for example, Bryant & Nunes, 2012; Denison & Xu, 2010; Girotto & Gonzalez, 2008; Reyna & Brainerd, 1994; Téglás et al., 2011, 2007). At the same time, in line with the results of this study, it has been shown (e.g., Bryant & Nunes, 2012; Garfield & Ahlgren, 1988) that children (but also most adults) find basic laws of probability difficult to grasp and apply.In order to reach a unified perspective which could provide, among other things, a solid scientific basis for effective teaching programs,future research should propose models able to explain both rea- soning biases and achievements. Among the latter are the good performances observed when adults can ground their probability judgments on intuitive causal models (Krynski & Tenenbaum, 2007; Meder, Mayrhofer, & Waldmann, 2014) or evidential impact evaluations (Mastropasqua, Crupi, & Tentori, 2010; Tentori, Chater, & Crupi, 2015; Tentori, Crupi, Bonini, & Osherson, 2007). In this regard, it would be interesting to see whether the same results hold with children, and whether their probability errors can be explained by their causal knowledge and/or by a prevalence of evidence assessment over probability appraisal (Tentori, Crupi, & Russo, 2013). What is clear, at the moment, is that the cognitive demands of understanding and dealing with probability play an important role but are not by themselves enough to account for children's reasoning failures and successes. To gain full understanding of both these phenomena, it is crucial to go beyond the focus on the timing of each specific capacity, and try to explain how reasoning competences are connected with and support each other. As stressed by Bryant and Nunes (2012), an important step in this direction could come from longitudinal studies that mea- sure how children's different reasoning competencies predict how easily they will learn later on. The ultimate aim is not justto draw up a list of what children can or cannot do at each age,but to provide a more meaningful picture including also the role of various reasoning processes in overall cognitive development.

Acknowledgements

Financial support for this study was provided by the Cassa di Risparmio of Trento and Rovereto (SMC, grant 40102595) and by the DFG as part of the priority program New Frameworks of Rationality (SPP 1516, grant CR 409/1-2). The study was approved by the Ethics Committee of CERME (University Ca' Foscari of Venice).

Appendix A

The ten Bayesian inference problems from Zhu and Gigerenzer (2006) that were employed in our pilot study. Correct responses are indicated within square brackets, while US adults' accuracy rates are reported in the right column.

	Problem	Accuracy rates (%)
1.	Pingping goes to a small village to ask for directions. In this village, 10 out of every 100 people will lie. Of the 10 people who lie, 8 have a red nose. Of the remaining 90 people who don't lie, 9 also have a red nose. Imagine that Pingping meets a group of people in the village with red noses. How many of these people will lie? [8 out of 17]	37.5
2.	There is a large package of sweet or salty cookies with various kinds of shapes. In the package, 20 out of every 100 cookies are salty. Of the 20 salty cookies, 14 are round. Of the remaining 80 sweet cookies, 24 are also round. Imagine you take out a pile of round cookies. How many of them are salty cookies? [14 out of 38]	22.5
3.	The principal of a school announced and explained a new school rule to all the students gathering together on the playground. Then the principal said: "Those who understand what I mean, please put up your hands." 70 out of every 100 students understood. Of these 70 who understood, 63 put up their hands. Of the remaining 30 who didn't understand, 9 put up their hands. Imagine a group of students who put up their hands. How many of them understood the principal? [63 out of 72]	20

(*continued on next page*)

Appendix A (*continued*)

	Problem	Accuracy rates (%)
4.	20 out of every 100 children in a school have bad teeth. Of these 20 children who have bad teeth, 10 love to eat sweet food. Of the remaining 80 children who don't have bad teeth, 24 also like to eat sweet food. Here is a group of children from this school who love to eat sweet food. How many of them may have bad teeth? [10 out of 34]	17.5
5.	To protect their children's eyes, mothers always urge children not to watch too much TV. Suppose you want to test this belief and get the following information: 30 out of every 100 children become near-sighted. Of these 30 near-sighted children, 21 of them watch too much TV. Of those 70 children with normal sight, 28 of them watch too much TV. Suppose you meet a group of children who watch too much TV, how many of them may become near-sighted? [21 out of 49]	25
6.	In Dongdong's town, 10 out of every 100 children are overweight. Of the 10 overweight children, 3 of them have overweight mothers. Of the remaining 90 children who have normal weight, 18 of them still have overweight mothers. Suppose you meet a group of overweight mothers in the town. How many of them have overweight children? [3 out of 21]	25
7.	A group of children are playing games with cards. Those who get a card with a picture of a cat on the inner side win a piece of candy. 30 of every 100 cards have a cat picture on one side. Of the 30 cards with a cat picture, 12 of them are red on the other side. Of the remaining 70 cards that have no cat pictures, 35 of them are still red on the other side. Imagine Dingding takes out a group of red cards. How many of them have a cat picture on the other side? [12 out of 47]	32.5
8.	In a cold winter in a town, 40 out of every 100 people hurt their hands by the cold. Of the 40 people who hurt their hands, 36 wear gloves in the open air. Of the remaining 60 people with normal hands, 30 also wear gloves. Suppose you meet a group of people who wear gloves in the town. How many of them hurt their hands? [36 out of 66]	30
9.	In a hospital, 60 out of every 100 patients get a cold. Of the 60 patients who get a cold, 42 have a headache. Of the remaining 40 patients with other diseases, 12 also have a headache. Suppose you meet a group of patients who have a headache in a hospital. How many of them get a cold? [42 out of 54]	30
10.	On a campus, 90 out of every 100 young people you meet are college students of this university. Of the 90 college students, 45 wear glasses. Of the remaining 10 young people that are not students of the university, 3 also wear glasses. Suppose you meet a group of young people who wear glasses on the campus. How many of them are students at this university? [45 out of 48]	32.5

Appendix B

The four problems and the corresponding questions employed in Experiment 2 in the natural frequency (left column) and in the number of chances groups (right column), adapted from Italian. Correct responses are reported within square brackets.

Natural frequency	Number of chances
Problem 1	
You went to a small village to ask for information.	You go to a small village to ask for information.
You met 100 persons in this village, 24 of them had a red nose.	When you meet a person in this village, there are 24 chances out of 100 that the person has a red nose.
15 of those 24 persons were liars.	In 15 of these 24 chances, the person is a liar.
The remaining 9 persons were not liars.	In the remaining 9 chances, the person is not a liar.
Imagine that you go back to this village and meet a new group of persons with red noses.	Imagine that you go back to this village and meet a new person with a red nose.
(*Distributive* question)	(*Distributive* question)
Out of every 24 persons with a red nose in this new group, there will be: _[15]_ persons who are liars and _[9]_ persons who are not liars.	For this new person with a red nose, there are: _[15]_ chances that s/he is a liar and _[9]_ chances that s/he is not a liar.
(*Standard* question)	(*Standard* question)
Thus, in this new group of persons with red noses, there will be: _[15]_ persons who are liars out of _[24]_	Thus, for this new person with a red nose, there are: _[15]_ chances that s/he is a liar out of _[24]_
(*Relative* question)	(*Relative* question)
Therefore, in this new group of persons with red noses, there will be more: ħ persons who are liars [x] ħ persons who are not liars	Therefore, for this new person with a red nose, there are more: ħ chances that s/he is a liar [x] ħ chances that s/he is not a liar

Appendix B (*continued*)

Natural frequency	Number of chances
Problem 2	
You went to a park to play. You met 100 persons in that park, 23 of them were wearing glasses.	You go to a park to play. When you meet a person in that park, there are 23 chances out of 100 that the person is wearing glasses.
9 of these 23 persons were students. The remaining 14 persons were not students. Imagine that you go back to that park and you meet a new group of persons wearing glasses.	In 9 of these 23 chances, the person is a student. In the remaining 14 chances, the person is not a student. Imagine that you go back to that park and you meet a new person wearing glasses.
(*Distributive question*) For every 23 persons wearing glasses in this new group, there will be: _[9]_ persons who are students and _[14]_ persons who are not students	(*Distributive question*) For this new person wearing glasses, there are: _[9]_ chances that s/he is a student and _[14]_ chances that s/he is not a student
(*Standard question*) Thus, in this new group of persons wearing glasses, there will be: _[9]_ persons who are students out of _[23]_	(*Standard question*) Thus, for this new person wearing glasses, there are: _[9]_ chances out of _[23]_ that s/he is a student
(*Relative question*) Therefore, in this new group of persons wearing glasses, there will be more: h persons who are students h persons who are not students [x]	(*Relative question*) Therefore, for this new person wearing glasses, there are more: h chances that s/he is a student h chances that s/he is not a student [x]
Problem 3	
Paul bought a detector machine tuned to selectively detect gold. When this machine finds a gold coin, it emits a sound. However, sometimes the machine fails and it emits a sound even if the coin is not a gold coin.	
Paul examined 100 coins with the gold detector machine, one by one. The machine emitted a sound for 19 coins. For 4 of these 19 coins the machine correctly emitted a sound, i.e., the coins were gold coins and the machine emitted a sound. For the remaining 15 coins, the machine incorrectly emitted a sound, i.e., the coins were not gold coins but the machine emitted a sound anyway. Imagine that Paul examines a new group of coins with the machine and that for some of them the machine emits a sound.	When Paul examines a coin with the gold detector machine there are 19 chances out of 100 that the machine emits a sound. In 4 out of these 19 chances the machine correctly emits a sound, i.e., the coin is a gold coin and the machine emits a sound. In the remaining 15 chances, the machine incorrectly emits a sound, i.e., the coin is not a gold coin but the machine emits a sound anyway. Imagine that Paul examines a new coin with the machine and that the machine emits a sound.
(*Distributive* question) Out of every 19 coins in this new group for which the machine emits a sound, there will be: _[4]_ gold coins for which the machine correctly emits a sound and _[15]_ coins that are not gold for which the machine emits a sound anyway.	(*Distributive question*) Given that the machine emits a sound for this new coin, there are: _[4]_ chances that it is a gold coin and the machine correctly emits a sound and _[15]_ chances that it is not a gold coin and the machine emits a sound anyway.
(*Standard* question) Thus, in this new group of coins for which the machine emits a sound, there will be: _[4]_ gold coins out of _[19]_	(*Standard* question) Thus, for this new coin for which the machine emits a sound there are: _[4]_ chances out of _[19]_ that it is a gold coin.
(*Relative* question) Therefore, in this new group of coins for which the machine emits a sound, there will be more: h gold coins h coins that are not gold coins [x]	(*Relative* question) Therefore, for this new coin for which the machine emits a sound, there are more: h chances that it is a gold coin h chances that it is not a gold coin [x]
Problem 4	
There is a new disease. A doctor has invented a machine to detect if a person has this disease. When a person is diseased, the machine emits a light. However, sometime the machine fails and it emits a light even if the person is not diseased.	
The doctor examined 100 persons with the machine, one by one. The machine emitted a light for 16 of them. For 13 of these 16 persons the machine correctly emitted the light, i.e., the person was diseased and the machine emitted a light. For the remaining 3 persons the machine incorrectly emitted a light, i.e., the person was not diseased but the machine emitted a light anyway.	Imagine that the doctor examines a new group of persons and thatfor some of them the machine emits a light.

When the doctor examines a person with the machine, there are 16 chances out of 100 that machine emits a light.
In 13 out of these 16 chances the machine correctly emits a light, i.e., the person is diseased and the machine emits a light.
In the remaining 3 chances, the machine incorrectly emits a light, i.e., the person is not diseased but the machine emits a light anyway.
Imagine that the doctor examines a new person, Anna, and that the machine emits a light.

(*Distributive* question)

Out of every 16 persons in this new group for which the machine emits a light, there will be: _[13]_ diseased persons for which

(*Distributive* question)

Given that the machine emits a light for Anna, there are: _[13]_ chances that Anna is diseased and the machine correctly emits a

(*continued on next page*)

Appendix B (*continued*)

Natural frequency	Number of chances
the machine correctly emits a light and _[3]_ persons that are not diseased for which the machine emits a light anyway.	light and _[3]_ chances that Anna is not diseased and the machine emits a light anyway.
(*Standard* question) Thus, in this new group of persons for which the machine emits a light, there will be: _[13]_ diseased persons out of _[16]_	(*Standard* question) Thus, for Anna, for whom the machine emits a light, there are: _[13]_ chances out of _[16]_ that she is diseased.
(*Relative* question) Therefore, in this new group of persons for which the machine emits a light, there will be more: ħ diseased persons [x] ħ persons that are not diseased	(*Relative* question) Therefore, for Anna, for whom the machine emits a light, there are more: ħ chances that she is diseased [x] ħ chances that she is not diseased

References

Acredolo, C., O'Connor, J., Banks, L., & Horobin, K. (1989). Children's ability to make probability estimates: Skills revealed through application of Anderson's Functional Measurement Methodology. *Child Development, 60*(4), 933–945.

Ansari, D. (2008). Effects of development and enculturation on numberrepresentation in the brain. *Nature Reviews Neuroscience, 9*(4), 278–291.

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *The Behavioral and Brain Sciences, 30*(3), 241–297.

Barth, H., La Mont, K., Lipton, J., Dehaene, S., Kanwisher, N., & Spelke, E. (2006). Non-symbolic arithmetic in adults and young children. *Cognition, 98*(3), 199–222.

Barth, H., La Mont, K., Lipton, J., & Spelke, E. S. (2005). Abstract number and arithmetic in preschool children. *Proceedings of the National Academy of Sciences of the United States of America, 102*(39), 14116–14121.

Beyth-Marom, R., & Fischhoff, B. (1983). Diagnosticity and pseudodiagnosticity. *Journal of Personality and Social Psychology, 45*(6), 1185–1195.

Brainerd, C. J. (1981). Working memory and the developmental analysis of probability judgment. *Psychological Review, 88*(6), 463–502.

Brainerd, C. J., & Reyna, V. F. (1990). Gist is the grist: Fuzzy trace theory and the new intuitionism. *Developmental Review, 10*(1), 3–47.

Bramwell, R., West, H., & Salmon, P. (2006). Health professionals' and service users' interpretation of screening test results: Experimental study. *BMJ: British Medical Journal, 333*(7562), 284.

Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychonomic Bulletin & Review, 15*(2), 284–289.

Brase, G. L., Pighin, S., & Tentori, L. (2017). (Yet) Another chance for good reasoning? A commentary and reply on Pighin, Tentori, and Girotto, 2017. *Psychonomic Bulletin & Review.* http://dx.doi.org/10.3758/s13423-017-1314-8 (in press).

Bryant, P., & Nunes, T. (2012). *Children's understanding of probability. A literature review* (summary report). Nuffield Foundation.

Butterworth, B., Reeve, R., Reynolds, F., & Lloyd, D. (2008). Numerical thought with and without words: Evidence from indigenous Australian children. *Proceedingsof the National Academy of Sciences of the United States of America, 105*(35), 13179–13184.

Campbell, J. I. D., & Xue, Q. (2001). Cognitive arithmetic across cultures. *Journal of Experimental Psychology: General, 130*(2), 299–315.

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), 1–73.

Cuite, C. L., Weinstein, N. D., Emmons, K., & Colditz, G. (2008). A test of numeric formats for communicating risk probabilities. *Medical Decision Making, 28*(3), 377–384.

Denison, S., & Xu, F. (2010). Twelve- to 14-month-old infants can predict single- event probability with large set sizes. *Developmental Science, 13*(5), 798–803.

Denison, S., & Xu, F. (2014). The origins of probabilistic inference in human infants. *Cognition, 130*(3), 335–347.

Evans, J. S. B. T., Handley, S. J., Perham, N., Over, D. E., & Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. *Cognition, 77* (3), 197–213.

Fischbein, E., & Gazit, A. (1984). Does the teaching of probability improve probabilistic intuitions? – An exploratory research study. *Educational Studiesin Mathematics, 15*(1), 1–24.

Fontanari, L., Gonzalez, M., Vallortigara, G., & Girotto, V. (2014). Probabilistic cognition in two indigenous Mayan groups. *Proceedings of the National Academy of Sciences of the United States of America, 111*(48), 17075–17080.

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

Fujimura, N. (2001). Facilitating children's proportional reasoning: A model of reasoning processes and effects of intervention on strategy change. *Journal of Educational Psychology, 93*(3), 589–603.

Galesic, M., Gigerenzer, G., & Straubinger, N. (2009). Natural frequencies help older adults and people with low numeracy to evaluate medical screening tests. *Medical Decision Making, 29*(3), 368–371.

Garcia-Retamero, R., & Hoffrage, U. (2013). Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Social Science & Medicine, 83*, 27–33.

Garfield, J., & Ahlgren, A. (1988). Difficulties in leaning basic concepts in probability and statistics – Implications for research. *Journal for Research in Mathematics Education, 19*(1), 44–63.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond ''heuristics and biases''. *European Review of Social Psychology, 2*(1), 83–115.

Gigerenzer, G. (1996). The psychology of good judgment. *Medical Decision Making, 16*(3), 273–280.

Gigerenzer, G. (2003). *Reckoning with risk: Learning to live with uncertainty*. UK: Penguin.

Gigerenzer, G. (2008). *Rationality for mortals.* New York: Oxford University Press.

Gigerenzer, G. (2015). On the supposed evidence for libertarian paternalism. *Review of Philosophy and Psychology, 6*(3), 361–383.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2008). Helping doctors and patients make sense of health statistics: Toward an evidence-based society. *Icots, 8*(2), 53–96.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*(4), 684–704.

Gigerenzer, G., & Hoffrage, U. (1999). Overcoming difficulties in Bayesian reasoning: A reply to Lewis and Keren (1999) and Mellers and McGraw (1999). *Psychological Review, 106*(2), 425–430.

Gigerenzer, G., & Hoffrage, U. (2007). The role of representation in Bayesian reasoning: Correcting common misconceptions. *Behavioral and Brain Sciences, 30*(3), 264–267.

Gigerenzer, G., Todd, P. M., & ABC Research Group, T. (1999). *Simple heuristics that make us smart.* Oxford University Press.

Girotto, V., & Gonzalez, M. (2001). Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition, 78*(3), 247– 276.

Girotto, V., Fontanari, L., Gonzalez, M., Vallortigara, G., & Blaye, A. (2016). Young children do not succeed in choice tasks that imply evaluating chances. *Cognition, 152*, 32–39.

Girotto, V., & Gonzalez, M. (2002). Chances and frequencies in probabilistic reasoning: Rejoinder to Hoffrage, Gigerenzer, Krauss, and Martignon. *Cognition, 84*(3), 353–359.

Girotto, V., & Gonzalez, M. (2008). Children's understanding of posterior probability. *Cognition, 106*(1), 325–344.

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive Science, 35*(8), 1407–1455.

Hayes, B. K., Hawkins, G. E., Newell, B. R., Hayes, B. K., Hawkins, G. E., & Newell, B. R. (2015). Consider the alternative: The effects of causal knowledge on representing and using alternative hypotheses in judgments under uncertainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(6), 723–739.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal, 327*(7414), 557–560.

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine: Journal of the Association of American Medical Colleges, 73*(5), 538–540.

Hoffrage, U., & Gigerenzer, G. (2004). How to improve the diagnostic inferences of medical experts. *Experts in Science and Society*, 249–268 (July 2015).

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition, 84*(3), 343–352.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J. P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review, 106*(1), 62–88.

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: Springer.

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General, 136*(3), 430–450.

Low, R., & Over, R. (1993). Gender differences in solution of algebraic word- problems containing irrelevant information. *Journal of Educational Psychology, 85*(2), 331–339.

Luecking, A. (2004). *The development of Bayesian reasoning in children* Unpublished diploma thesis. Berlin: Free University.

Macchi, L. (1995). Pragmatic aspects of the base-rate fallacy. *The Quarterly Journal of Experimental Psychology, 48*(1), 188–207.

Martignon, L., & Kuntze, S. (2015). Good models and good representations area support for learners' risk assessment. *The Mathematics Enthusiast, 12*, 157–167.

Marzocchi, G. M., Lucangeli, D., De Meo, T., Fini, F., & Comoldi, C. (2002). The disturbing effect of irrelevant information on arithmetic problem solving in inattentive children. *Developmental Neuropsychology, 21*(1), 73–92.

Mastropasqua, T., Crupi, V., & Tentori, K. (2010). Broadening the study of inductive reasoning: Confirmation judgments with uncertain evidence. *Memory & Cognition, 38*(7), 941–950.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review, 121*(3), 277–301.

Multmeier, J. (2012). Representations facilitate Bayesian reasoning: computational facilitation and ecological design revisited. Unpublished manuscript.

Noelting, G. (1980). The development of proportional reasoning and the ratio concept. Part I – Differentiation of stages. *Educational Studies in Mathematics, 11* (2), 217–253.

Nunes, T., & Bryant, P. (1996). *Children doing mathematics*. Wiley-Blackwell.

OECD (2015). *PISA 2015 results (Volume I). Excellence and equity in education*. http://dx.doi.org/10.1787/9789264266490-en. Retrieved from <https://www.oecd.org/pisa/>.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children* (L. Leake et al., Trans.). Oxford, England: W.W. Norton (p. 251).

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Materials and Method. Science, 306*(5695), 1–16.

Pighin, S., Tentori, K., Savadori, L., & Girotto, V. (2017). Fostering the understanding of positive test results (submitted for publication).

Pighin, S., Tentori, K., & Girotto, V. (2017). Another chance for good reasoning. *Psychonomic Bulletin & Review*. http://dx.doi.org/10.3758/s13423-017-1252-5 (in press).

Pighin, S., Gonzalez, M., Savadori, L., & Girotto, V. (2015). Improving public interpretation of probabilistic test results: Distributive evaluations. *Medical Decision Making, 35*(1), 12–15.

Pighin, S., Gonzalez, M., Savadori, L., & Girotto, V. (2016). Natural frequencies do not foster public understanding of medical test results. *Medical Decision Making, 36*, 686–691.

Pitkethly, A., & Hunting, R. (1996). A review of recent research in the area of initial fraction concepts. *Educational Studies in Mathematics, 30*(1), 5–38.

Posner, M. I., & Rothbart, M. K. (2005). Influencing brain networks: Implications for education. *Trends in Cognitive Sciences, 9*(3), 99–103.

Rakoczy, H., Cluever, A., Saucke, L., Stoffregen, N., Graebener, A., Migura, J., & Call, J. (2014). Apes are intuitive statisticians. *Cognition, 131*(1), 60–68.

Reyna, V. F. (1991). Class inclusion, the conjunction fallacy, and other cognitive illusions. *Developmental Review, 11*(4), 317–336.

Reyna, V. F. (2004). How people make decisions that involve risk. A dual-processes approach. *Current Directions in Psychological Science, 13*(2), 60–66.

Reyna, V. F., & Brainerd, C. J. (1995a). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences, 7*(1), 1–75.

Reyna, V. F., & Brainerd, J. (1995b). Interim theory: Synthesis. *Learning and Individual Differences, 7*(1), 1–75.

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences, 18*(1), 89–107.

Reyna, V. F., & Brainerd, C. J. (1994). The origins of probability judgment: A review of data and theories. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 239–270). Oxford, England: John Wiley & Sons.

Schwartz, D. L., & Moore, J. L. (1998). On the role of mathematics in explaining the material world: Mental models for proportional reasoning. *Cognitive Science, 22* (4), 471–516.

Skov, R. B., & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology, 22*(2), 93–121.

Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions andother fallacies. *Organizational Behavior and Human Decision Processes, 91*(2), 296–309.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28*(3), 303–333.

Spinillo, A. G., & Bryant, P. (1991). Children's proportional judgments: The importance of ''half''. *Child Development, 62*(3), 427–440.

Tang, Y., Zhang, W., Chen, K., Feng, S., Ji, Y., Shen, J., ... Liu, Y. (2006). Arithmetic processing in the brain shaped by cultures. *Proceedings of the National Academy of Sciences, 103*(28), 10775–10780.

Téglás, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences of the United States of America, 104*(48), 19156–19159.

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science, 332* (6033), 1054–1059.

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

Tentori, K., Chater, N., & Crupi, V. (2015). Judging the probability of hypotheses versus the impact of evidence: Which form of inductive inference is more accurate and time-consistent? *Cognitive Science, 40*(3), 758–778.

Tentori, K., Crupi, V., Bonini, N., & Osherson, D. (2007). Comparison of confirmation measures. *Cognition, 103*(1), 107–119.

Tentori, K., Crupi, V., & Russo, S. (2013). On the determinants of the conjunction fallacy: Probability versus inductive confirmation. *Journal of Experimental Psychology: General, 142*(1), 235–255.

Torbeyns, J., Schneider, M., Xin, Z., & Siegler, R. S. (2015). Bridging the gap: Fraction understanding is central to mathematics achievement in students from three different continents. *Learning and Instruction, 37*, 5–13.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293.

Waters, E. A., Weinstein, N. D., Colditz, G. A., & Emmons, K. (2006). Formats for improving risk communication in medical tradeoff decisions. *Journal of Health Communication, 11*(2), 167–182.

Yost, P. A., Siegel, A. E., & Andrews, J. M. (1962). Nonverbal probability judgments by young children. *Child Development, 33*(4), 769–780.

Zhu, L., & Gigerenzer, G. (2006). Children can solve Bayesian problems: The role of representation in mental computation. *Cognition, 98*(3), 287–308.