



This postprint was originally published by the American Psychological Association as:

Gigerenzer, G., Multmeier, J., Föhring, A., & Wegwarth, O. (2021).

Do children have Bayesian intuitions? *Journal of Experimental Psychology: General*, 150(6), 1041–1070.

<https://doi.org/10.1037/xge0000979>

The following copyright notice is a publisher requirement:

©American Psychological Association, 2021. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. Please do not copy or cite without author's permission.

The final article is available, upon publication, at:

<https://doi.org/10.1037/xge0000979>

Provided by:

Max Planck Institute for Human Development

Library and Research Information

library@mpib-berlin.mpg.de

Do Children Have Bayesian Intuitions?

Gerd Gigerenzer, Jan Multmeier, Andrea Föhrling, and Odette Wegwarth
Max Planck Institute for Human Development, Berlin, Germany

Can children solve Bayesian problems, given that these pose great difficulties even for most adults? We present an ecological framework in which Bayesian intuitions emerge from a match between children's numerical competencies and external representations of numerosity. *Bayesian intuition* is defined here as the ability to determine the *exact* Bayesian posterior probability by minds untutored in probability theory or in Bayes' rule. As we show, Bayesian intuitions do not require processing of probabilities or Arabic numbers, but basically the ability to count tokens in icon arrays and to understand what to count. A series of experiments demonstrates for the first time that icon arrays elicited Bayesian intuitions in children as young as second-graders for 22% to 32% of all problems; fourth-graders achieved 50% to 60%. Most surprisingly, icon arrays elicited Bayesian intuitions in children with dyscalculia, a specific learning disorder that has been attributed to genetic causes. These children could solve an impressive 50% of Bayesian problems, a level similar to that of children without dyscalculia. By seventh grade, children solved about two thirds of Bayesian problems with natural frequencies alone, without the additional help of icon arrays. We identify four non-Bayesian rules. On the basis of these results, we propose a common solution for the phylogenetic, the ontogenetic, and the 1970s puzzles in the Bayesian literature and argue for a revision of how to teach statistical thinking. In accordance with recent work on infants' numerical abilities, these findings indicate that children have more numerical ability than previously assumed.

Keywords: Bayesian intuition, dyscalculia, icon array, natural frequencies, cognitive development

Theoretical concepts studied in psychology typically have a long history that can be traced back to Aristotle and earlier—such as association, memory, and causality. Mathematical probability is one exception to this rule, uncovered so late that Hacking (1975) referred to it as a “scandal of philosophy.” He dated its discovery to 1654, in an exchange of letters about gambling problems between the French mathematicians Blaise Pascal and Pierre Fermat. Their work solved the problem of *direct probability*, which asks: If the proportion of balls of different colors in an urn is known a priori, determine the probable results of random drawings. It took another century before a solution was found to the problem of *inverse probability*: If only the results of the drawings are known, determine the probable mixture of balls in the urn. The direct problem was originally framed as an inference from cause to effect, and the inverse problem as one from effect to cause (Daston, 1988). In modern terminology, the inverse problem entails inferring the probability of a hypothesis given data.

The solution to the inverse problem has been named *Bayes' rule*, based on an essay by the Nonconformist minister Thomas Bayes that was edited and published posthumously in 1763 by his friend Richard Price (Bayes, 1763). To Price, the rule provided a proof of the existence of God, the ultimate cause (Stigler, 1999). Bayes' essay had no discernible impact and sank from view almost immediately. If the mathematician Pierre Laplace (1781) had not independently rediscovered Bayes' rule, it might have ended up as a forgotten treasure.

The study of probabilistic reasoning in psychology has a similarly belated history. It is strikingly absent before 1950 (Gigerenzer & Murray, 2015). The question of whether the mind is an intuitive statistician—Bayesian or otherwise—virtually never occurred to psychologists who studied thinking and reasoning before the mid-20th century. Probability was about judgmental error or stochastic reinforcement (Brunswik, 1939), not the laws of thought. The first experiments on adults' Bayesian reasoning were conducted by Rouanet (1961) in France and by Edwards (1968) in the United States. Eventually, in the wake of the cognitive revolution, various statistical methods for inference, including those of Bayes, Fisher, and Neyman–Pearson, turned into theories of cognition (Gigerenzer, 1992), as did the computer (Gigerenzer & Goldstein, 1996). During the cognitive revolution, the mind made its comeback, shouldered by statistical tools.

The study of the concept of probability in children also has a delayed history. Its beginnings can be dated to Piaget and Inhelder's *The Origin of the Idea of Chance in Children*, first published in 1951 in French. It took more than 20 years before the book was published in English, namely in 1975. Piaget and Inhelder studied notions of irreversibility, the law of large numbers, and other concepts related to direct probability, referencing the work of mathematicians Richard von Mises and Hans Reichenbach, among others. Bayes' rule, in contrast, is not mentioned. In

 [Gerd Gigerenzer](#), Jan Multmeier, Andrea Föhrling, and Odette Wegwarth, Max Planck Institute for Human Development, Berlin, Germany.

This research was supported by the Max Planck Society. Gerd Gigerenzer has presented the basic idea and part of Study 1 in a lecture to the Royal Academy of Arts and Sciences, Amsterdam, July 2014, and in a brief popular account in *Risk Savvy* (Gigerenzer, 2014, pp. 248–250).

Correspondence concerning this article should be addressed to Gerd Gigerenzer, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: gigerenzer@mpib-berlin.mpg.de

The Intuitive Sources of Probabilistic Thinking in Children, Fischbein distinguished two lines of research: probability learning, following Brunswik, and the development of concepts such as chance and proportion, following Piaget (Fischbein 1975). Again, no mention is made of Bayes' rule. Bayes' rule is also absent in Gelman and Gallistel's (1978) influential *The Children's Understanding of Number*, in Dehaene's (1997/2011) *The Number Sense*, and in overviews on numerical development (Siegler & Braithwaite, 2017).

Two Bayesian approaches have emerged in recent years: *Bayesian modeling* and *Bayesian problem-solving*. The first aims to model children's performance in tasks such as learning concepts and acquiring language within a Bayesian framework (e.g., Tenenbaum, Kemp, Griffiths, & Goodman, 2011). For instance, in one study children were told that a toy object is a "blink" (a previously unknown concept) and that their task was to pick other "blinks" from a selection of toys. Children's pattern of generalization could be fitted better by a Bayesian model than by competing mathematical models (Xu & Tenenbaum, 2007). This approach builds Bayesian models at the level of computational theory but makes no claim that children in fact are consciously or unconsciously able to perform the computations (Xu & Tenenbaum, 2007, p. 270). The second approach, Bayesian problem solving, directly addresses this question and has its origin in the earlier research mentioned above on adults by Rouanet and Edwards. Here, people are given quantitative Bayesian reasoning problems and are directly tested whether they can estimate the posterior probability. This article deals with—and extends—the Bayesian problem-solving approach. Two kinds of Bayesian problem-solving tasks have been distinguished in the literature: The first involves *judgments of proportions*, and the other *judgments of relative likelihood* (Gallistel, Krishan, Liu, Miller, & Latham, 2014; Peterson & Beach, 1967). Here we deal exclusively with the former kind, where a proportion " x out of y " is estimated (see below). In judgments of relative likelihood (such as whether a sample has been drawn from urn A with 30 red and 70 black balls, or from urn B with 70 red and 30 black balls), by contrast, the proportions are known and the relative likelihood of the hypotheses (A or B) is estimated. Both approaches are complementary and address different questions. To solve proportion problems is extremely challenging; A meta-analysis reported that only 4% of adults could find the Bayesian posterior probability (McDowell & Jacobs, 2017).

By the end of the 20th century, no single study had investigated whether children are able to make judgments according to Bayes' rule. To the best of our knowledge, the first study appeared in 2006, with Chinese children (Zhu & Gigerenzer, 2006). This striking delay is possibly attributable to the widely accepted view in late-20th-century psychology that even adults are largely unable to solve Bayesian problems. According to influential researchers, people are "not Bayesian at all" (Kahneman & Tversky, 1972, p. 450), and "fail to make forecasts that are consistent with Bayes' rule" (Thaler & Sunstein, 2003, p. 176). If adults have no Bayesian intuitions, then there is little reason to believe that children do.

In the past decades, research on the "number sense" discovered that babies and young children are better intuitive statisticians than previously assumed by Piaget and others (Carey, 2009; Gelman & Gallistel, 1978). As a consequence, Gopnik (2014) wondered why babies are so adept when adults are often so incompetent in dealing with numbers. In this article we show that, like infants' number sense, children's ability to reason the Bayesian way has been left unrecognized. Moreover, in the General Discussion section, we use these results to provide an explanation for the apparently contradictory findings about children's and adults' ability to reason the Bayesian way.

We first present an ecological theory of Bayesian intuitions that integrates the developmental approach to numeracy with the theory of ecological rationality (Gigerenzer, Hertwig, & Pachur, 2011). An ecological approach explains behavior as a function of cognitive abilities *and* the external representation of numerosity, not of internal processes alone, as in Piaget and Inhelder (1951/ 1975). To do justice to the numerical competencies of young schoolchildren, we extend earlier work on natural frequencies (Gigerenzer & Hoffrage, 1995, 1999) to nonnumerical representations called *icon arrays*. We show that Bayesian intuitions require neither understanding nor processing of mathematical probabilities or of Arabic numerals, but basically only the ability to count, such as tokens in icon arrays, and to know what to count and how to combine the counts. In a series of studies, we then test the prediction that icon arrays can elicit exact Bayesian intuitions in children.

An Ecological Theory of Bayesian Intuitions

We use the term *intuition* to refer to judgments of children (or adults) who have had no instruction in probability theory or in Bayes' rule. We call an intuition a *Bayesian intuition* if it corresponds *exactly* to the Bayesian posterior probability. We use the term *correspond* here because children will be asked for frequency estimates—such as x out of y , where x and y are natural numbers—in place of probabilities, given that they do not yet understand the latter, which are decimals. Thus, when referring to estimates of the Bayesian posterior probability in this article, we operationalize these with equivalent frequency estimates. The subject of this article is hence the untutored mind.

We use *internal* versus *ecological* to distinguish two kinds of theories. Internal theories view behavior as a function of internal rules plus noise. For instance, virtually all explanations of adults' correct or incorrect Bayesian reasoning have been internal, such as postulating approximate Bayesian processes that overweigh base rates (conservatism; Edwards, 1968), or, by contrast, genuinely non-Bayesian processes such as the representativeness heuristic and base rate neglect (Tversky & Kahneman, 1980). These theories share the property that their concepts consist exclusively of internal processes. Piaget and Inhelder's (1951/1975) stage theory also postulates a progression of internal operations: a prelogical stage characterized by the absence of operations such as hierarchic nesting of sets, followed by the emergence of these operations but limited to concrete objects that can be seen or handled, and, finally, the emergence of formal thought, that is, probabilistic reasoning independent of the concrete content.

In contrast, an ecological theory of Bayesian intuition views behavior as a function of both cognition and environment. Ecological views have been emphasized at many points in time. According to Lewin's (1936) field theory, behavior is a function of person *and* environment; in Brunswik's (1955) probabilistic functionalism, cognition and environment have to come to terms by mutual adaptation; and in Simon's (1990) scissors analogy, behav-

ior is a function of the match between two blades, cognition *and* environment. An ecological theory of Bayesian intuitions investigates the match between the external representation of numerosity and the internal numerical abilities of the child.

External Representation of Numerosity

Numerosity is the physical quantity of a set, distinct from its representation by humans such as in the form of a tally or an Arabic numeral. Representations, often called *notations*, play a decisive role in the development of mathematics. In the history of numbers, the basic notations were *concrete*: a tally of similar tokens (e.g., fingers, notches on a stick) that represent a one-to-one mapping of a physical set (Ifrah, 2000). For instance, in Egyptian hieroglyphic notation, *four* is written as *IIII*. The cardinal system of the Romans replaced the earlier picturesque symbols of the Egyptians and provided a better solution for representing large numerosities. At some time in the first centuries A.D., an unknown Hindu scholar made one of the most important inventions in mathematics, *positional notation*, with a zero as a place holder, which was later adopted by Arab mathematicians in the 10th century and is since known as the Arabic numeral system (Dantzig, 1954, p. 31). In Europe, the struggle between the old notations and the positional system went on for centuries, inhibiting progress, until finally the “algorists” (who used the positional system) won over the “abacists” (Dantzig, 1954). At the end of his *Universal History of Numbers*; Ifrah (2000) poses the question whether modern mathematics, with all its practical applications that have revolutionized our lives, could have possibly occurred in the absence of a positional numerical notation. His answer is: “It seems incredibly unlikely.”

The general point is that a specific representation of numerosity is not neutral but can facilitate or hinder calculation and understanding. If the Romans had difficulties doing division, then the cause was located not simply in their minds, but also in the notation. The Arab system facilitates multiplication and division, which Roman numerals clearly do not.

A similar argument can be made with respect to the delayed acceptance of Bayes’ original paper. Bayes (1763) introduced his argument with a peculiar geometric mode of description used earlier by Newton, together with a physical mechanism: balls thrown onto the surface of a table. This choice of representation may have been a factor in the subsequent neglect of his discovery. In fact, what we today call Bayes’ rule cannot be found in Bayes’ original paper. In contrast, Laplace’s (1781) representations—such as drawing tickets from urns—proved to be more successful and are what we today associate with Bayes’ rule and conveniently use to test Bayesian reasoning (Stigler, 1986).

We argue that external representations are equally crucial for Bayesian intuitions. The general argument is:

1. Every numerosity has multiple representations.
2. Different representations can require different numerical competencies for calculating the Bayesian posterior probability.
3. If we match the representation to the competencies of children of a particular age, we should be able to elicit Bayesian intuitions, if these indeed exist in children.

Development of Numerical Competencies

There is wide agreement that human newborns are able to distinguish small numerosities (Antell & Keating, 1983; Wynn, 1992; Xu & Spelke, 2000), an ability that has been referred to as the *number sense* (Dantzig, 1954). Dehaene (1997/2011) proposed the *number sense hypothesis*, the idea that we owe our mathematical capacities to an inherited capacity shared with other animals, that is, the rapid perception of approximate numbers of objects, possibly located in the intraparietal sulcus of both hemispheres, a brain system that is available to various animal species as well as to preverbal human infants. The quick grasp of numerosities up to four or five is called subitizing, but how it works is not well understood; Gelman and Gallistel (1978) conjectured that it is essentially just quick counting. The number sense can be impaired; children with dyscalculia at age 11, despite being schooled, are reported to have a number sense equivalent to that of 5-year-olds (Dehaene, 1997/2011, p. 267). In young children, numerosities can be determined by matching, such as by representing every token in a set by a notch on a stick. This principle of one-to-one *correspondence* between two sets does not yet require counting.

Counting means to assign to every token in a set an ordered sequence of number terms, such as one, two, three. It is based on correspondence and *succession* (Dantzig, 1954). Children as young as 4 to 5 years are reported to be able to count sets of five to 19 objects correctly, particularly if they have a minute of time to do so, but with large interindividual variability (Gelman & Tucker, 1975). Counting also requires understanding that the final number counted is the cardinal number of the set. Counting proceeds with natural numbers, which are the only numbers a young child recognizes (Gelman & Gallistel, 1978). In Western countries, most children are able to count, add, and subtract small numbers when they enter first grade (Starkey, Spelke, & Gelman, 1990).

The ability to understand Arabic numerals is acquired in numerate societies after counting has been mastered. To understand, add, and multiply Arabic numbers is a surprising ability, given that it took humankind millennia to develop this positional system with a zero place holder, and another millennium to introduce it in Europe. Nevertheless, 5- and 6-year-old kindergarten children who had not learned how to add and subtract already showed an approximate understanding of adding and subtracting two-digit numbers (Gilmore, McCarthy, & Spelke, 2007). Teaching animals symbolic notations of numbers, in contrast, requires much time and patience and has been mostly limited to one-digit numbers (Gelman & Gallistel, 1978, pp. 35–39).

Understanding probabilities proves to be a much more difficult task. Unlike natural numbers, probabilities are fractions between 0 and 1; alternatively, they are often represented as percentages between 0 and 100. Piaget and Inhelder (1951/1975) concluded that children master the principles of probability only when they reach the stage of formal operations, by age 11 to 12; in their review, Hoemann and Ross (1982) agreed, concluding that younger children are without even a “primitive conception of probability” (p. 116). In addition, many older children, despite having been taught fractions, have great difficulties understanding how to add or multiply these (Siegler & Braithwaite, 2017). The difficulty with adding fractions extends into adulthood, as illustrated by the so-called “freshman error” of adding numerators and adding denominators, such as $1/2 + 1/3 = 2/5$ (Silver, 1986).

Without understanding probabilities and how to add and multiply these, however, one cannot solve the typical Bayesian problems posed in psychological experiments (e.g., Tversky & Kahneman, 1980). Professionals such as physicians face similar problems. Consider a patient with a positive hemmocult screening test result who wants to know the actual likelihood of having colon cancer. In a study, 24 experienced physicians—including heads of medical departments—were given the same information: a base rate of 0.3%, a sensitivity of 50%, and a false positive rate of 3%. Their estimates of the probability of colon cancer given a positive test result, however, ranged between 1% and 99% (Hoffrage & Gigerenzer, 1998). Bayes' rule results in an estimate of about 5%, found by only one physician; the modal estimate was 50%. Many physicians also openly admit that they do not understand probabilities (Gigerenzer, 2014). The failure of psychology students and professionals alike has been widely interpreted as implying that the mind is not Bayesian. An ecological theory provides an alternative explanation: A negative conclusion is premature because participants were tested with a representation of numerosity that most of them do not master, which requires understanding, adding, and multiplying probabilities, including conditional probabilities.

Bayesian Intuition = Numerical Competence × Numerosity Representation

The central thesis of this article is that Bayesian intuitions are a function of the match between numerical competence and external representation of numerosity. Gigerenzer and Hoffrage (1995, 1999) showed that natural frequencies facilitate Bayesian reasoning relative to conditional probabilities. We extend this research to nonnumerical representations, specifically icon arrays, and provide a comparative analysis of the cognitive abilities required by all three representations: conditional probability, natural frequency, and icon array.

Bayesian Intuition and Conditional Probability

The Bayesian problems considered in this article correspond to those typically studied with adults (e.g., Brase, 2008; Hafenbrädl & Hoffrage, 2015; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000; McDowell & Jacobs, 2017). Each problem involves two binary variables, such as cause and effect, disease and symptom, or, more generally, hypothesis (H) and data (D). In this elementary case, Bayes' rule can be written as:

$$p(H|D) = p(H)p(D|H)/p(D) \quad (1)$$

where $p(H|D)$ is the posterior probability, that is, the probability of H given D ; $p(H)$ is the prior probability; $p(D|H)$ is the conditional probability of D given H ; and $p(D)$ is the probability of D .

To illustrate, consider the “magic wand problem” used in Studies 1, 2, and 3. In this problem, inspired by the Harry Potter stories, the desired object is a magic wand, and the task is to estimate the chance that students own a magic wand if they own a magic hat. First, we present the information in the *conditional probability* format, typically used in studies with adults. In this format, three pieces of information are given: the base rate (which serves as prior probability) and two conditional probabilities, the hit rate and the false alarm rate.

The Magic Wand Problem in Conditional Probabilities:

In Hogwarts School of Witchcraft, the probability that a student has a magic wand is 25%. If a student has a magic wand, the probability that the student has a magic hat is 80%. If a student does not have a magic wand, the probability that student also has a magic hat is 80%. Imagine you meet a student at Hogwarts School with a magic hat. What is the probability that the student has a magic wand? ____%

The solution of the Magic Wand Problem requires an extended version of Equation 1:

$$p(H|D) = p(H)p(D|H)/[p(H)p(D|H) + p(-H)p(D|-H)] \quad (2)$$

where H stands for “magic wand,” $-H$ for “no magic wand,” and D for “magic hat.” Inserting the numbers (as probabilities rather than percentages) into Equation 2, the probability $p(\text{magic wand}|\text{magic hat}) = (.25)(.80)/[(.25)(.80) + (.75)(.80)] = .25$. That is, the probability that a student with a magic hat actually has a magic wand is .25, or 25%.

When given conditional probabilities, many adults cannot find the Bayesian posterior probability, as has been shown in numerous studies with similar problems. As mentioned above, in their metaanalysis, McDowell and Jacobs (2017) estimated that only 4% of participants could find the Bayesian posterior probability. This failure has been attributed to internal factors, such as representativeness leading to base rate neglect (e.g., Kahneman, 2011; Tversky & Kahneman, 1980). Yet protocol analyses of failed solutions revealed that adults tried a variety of non-Bayesian strategies that do and do not ignore base rates, such as $p(H)p(D|H)$, which ignores $p(D|-H)$ but not base rates, and $p(D|H)$, which ignores both (Gigerenzer & Hoffrage, 1995, 2007). Rather than a systematic non-Bayesian rule, individual analysis of solutions suggest that most adults simply do not know how to solve these problems and randomly pick one or two of the numbers given, such as “50%” in the case of the physicians above.

What numerical competencies are required for Bayesian intuitions on the basis of conditional probabilities? We adopt Gelman and Gallistel's (1978) distinction between *number abstraction* and *number reasoning* processes. Number abstraction is the ability to abstract numbers from the information given, such as understanding Arabic or Roman numerals. Number reasoning is the ability to reason about the abstracted numbers, such as performing addition or multiplication. Based on this distinction, Equations 1 and 2 show that a probability representation requires the following cognitive competencies:

- P1. The ability to understand unconditional and conditional probabilities (which presupposes understanding fractions or decimals between 0 and 1, or percentages between 0 and 100).
- P2. The ability to identify the four relevant probabilities $p(H)$, $p(-H)$, $p(D|H)$, and $p(D|-H)$.
- P3. The ability to add, multiply, and divide probabilities.

These three competencies are necessary for Bayesian intuitions with conditional probabilities but not sufficient. In addition, Bayesian intuitions require that the probabilities are combined according to Equation 2. The competency P1 concerns the capability to abstract numbers from the information given, and P2 and

P3 are the reasoning processes required. Many adults have not mastered P1 to P3, and even P1 is difficult for some. For instance, in a representative sample of U.S. citizens, 30% could not convert “1%” to 10 out of 1,000, and 75% could not convert “1 in 1,000” into a percentage (Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007). The freshman error is a specific instance of the inability to do arithmetic with fractions (P3). The Berlin Numeracy Test, a quick test whose successful solution requires the abilities P1 to P3, has shown the limits of people’s abilities to reason with probabilities and percentages in 15 countries (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012).

Because we are dealing with children untutored in mathematical probability, P1 to P3 cannot be expected to be part of their repertoire of competencies. This expectation has been empirically demonstrated in a study with Chinese children who were tested on 10 problems similar to the magic wand problem: Not a single child could solve a single Bayesian problem (Zhu & Gigerenzer, 2006). To the best of our knowledge, this is the only existing study that tested children’s Bayesian intuitions given information presented in conditional probabilities (Figure 1).

A few studies have tested children’s understanding of base rates, such as asking whether it is more likely to draw a black token from a bag with predominantly white or black events (a relative likelihood task). In one of these studies, when numerical base rate information was given together with nonnumerical individuating information, attention to base rates was not found to increase from first to sixth grade, in both the object and social domains (Jacobs & Potenza, 1991, p. 170). Such results appear to confirm that children, like most adults, fail to have the cognitive preconditions for Bayesian intuitions. Others have tested whether children are able to qualitatively update judgments based on new evidence (Giroto & Gonzalez, 2008), but none of these studies required quantitative estimates of the Bayesian posterior probability.

Bayesian Intuition and Natural Frequency

The defining feature of conditional probabilities is conditionalization. For instance, $p(D|H)$ is conditionalized relative to H .

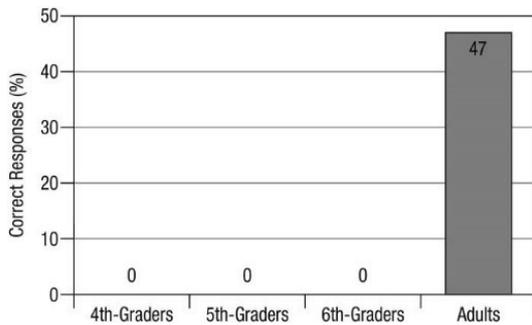


Figure 1. Percentage of Bayesian intuitions when the information is represented in conditional probabilities. Based on 60 participants (10 in each of the children groups and 30 adults). Children were from ordinary elementary schools in Beijing; adults were MBA students from the School of Management, Beijing University of Aeronautics and Astronautics. Each participant worked on 10 problems similar to the magic wand problem. None of the children could solve any of the problems. Adapted from Figure 1 of “Children can solve Bayesian problems: The role of representation in mental computation,” by L. Zhu and G. Gigerenzer, 2006, *Cognition*, 98, pp. 287–308. Copyright 2006 by Elsevier, Amsterdam, the Netherlands. Adapted with permission.

Unlike conditional probabilities, natural frequencies are not conditionalized relative to H or $\neg H$. The qualifier *natural* reflects this feature and distinguishes natural from relative frequencies, which, like conditional probabilities, are also conditionalized. Natural frequencies are a final tally of a process of updating frequencies called *natural sampling* (Kleiter, 1994). For illustration, consider a physician confronted with a new disease who has discovered a symptom that signals the disease. After having seen 100 patients, one by one, there were 20 cases with disease and symptom, five with disease but without symptom, 10 without disease but with symptom, and 65 without disease and without symptom. These four numbers are natural frequencies, that is, joint frequencies (such as disease *and* symptom), not simple frequencies. The physician can estimate the posterior probability $p(\text{disease}|\text{symptom})$ that a new patient with the symptom actually has the disease as $20/(20 + 10)$, or 67%. The physician learns natural frequencies from experience, while the conditional probability format corresponds to learning probabilities from textbooks. Natural sampling is a more general principle of learning; it is implemented in the way supervised neural networks learn as well as in human frequency learning through updating the frequency of joint events (Gallistel et al., 2014).

Figure 2 shows the difference between conditional probability and natural frequency representations using a tree format of the magic wand problem. The tree to the left is a conditional probability tree, with the prior probabilities (base rates) on the middle level, and the four conditional probabilities on the bottom level. Conditionalization occurs at the transition from the middle to the bottom level. Through conditionalization, all information about base rates is lost.

In the natural frequency tree, the absolute numbers are kept relatively small because this and similar problems in Studies 1, 2, and 3 were used to test second- through fourth-graders. The four numbers at the bottom are the natural frequencies (a to d). There is no conditionalization, which can be seen from the *additivity principle*: These four natural frequencies add up to the number at the top of the tree. That means that one can add up “4 of 5” and “12 of 15” to equal “16 of 20,” which appears to be the freshman’s error but is in fact a correct operation with natural frequencies. The additivity principle does not hold in a conditional probability tree or a relative frequency tree. With this analysis, we can now reformulate the magic wand problem.

The Magic Wand Problem in Natural Frequencies:

Out of every 20 students of Hogwarts School of Witchcraft, five have a magic wand. Of these five students, four also have a magic hat. Of the other 15 students without a magic wand, 12 also have a magic hat. Imagine you meet a group of students at Hogwarts School who have a magic hat. How many of them have a magic wand? ___ out of ___.

Conditional probability and natural frequencies are mathematically equivalent in the sense that natural frequencies can be mapped one-to-one into conditional probabilities, and vice versa (except for the choice of the sample size, here 20). Natural frequencies also share key properties with probabilities, such as closure, commutativity, and associativity of their addition (Weber, Binder, & Krauss, 2018).

However, although these representations are mathematically equivalent, they are not computationally and psychologically

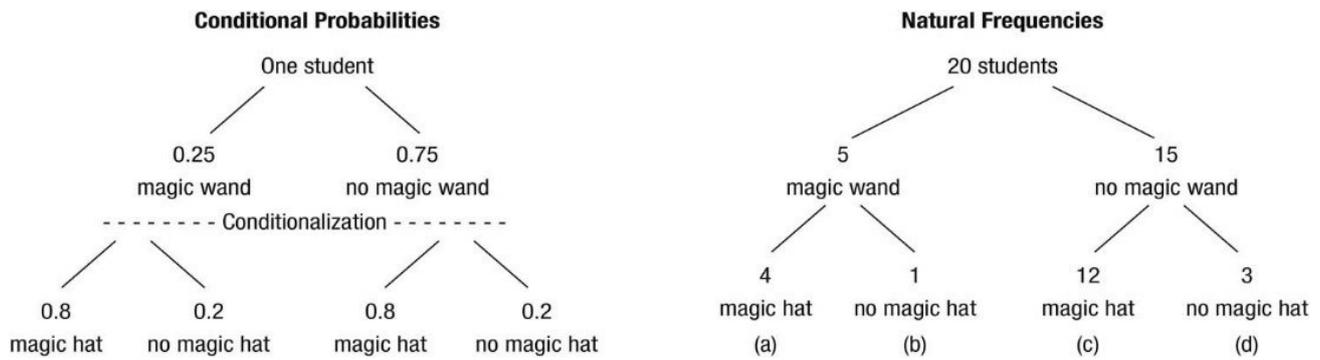


Figure 2. Conditional probability (left) and natural frequency representation (right) of the magic wand problem in a tree form. The four conditional probabilities at the bottom of the left-hand tree do not contain any information about base rates; the four natural frequencies at the bottom of the right-hand tree do. In the middle layer of the left-hand tree, the two probabilities are the base rate probabilities (which are not conditional probabilities), and in the middle layer of the right-hand tree, the two frequencies are the base rate frequencies (which are not natural frequencies). Conditionalization occurs between the bottom two levels in the conditional probability tree; it does not occur in the natural frequency tree. A tree with relative frequencies is identical to the conditional probability tree.

equivalent. In what follows, we show that natural frequencies reduce Bayesian computations radically. Bayes' rule in the form of Equation 2 is:

$$p(H|D) = p(H)p(D|H) / [p(H)p(D|H) + p(-H)p(D|-H)].$$

We now replace the probabilities with natural frequencies, using the symbols a , b , c , and d for the four natural frequencies in Figure 2 (right; Kleiter, 1994). We get:

$$p(H|D) = \frac{a+b}{a+b+c+d} \frac{a}{a+b} / \left[\frac{a+b}{a+b+c+d} \frac{a}{a+b} + \frac{c+d}{a+b+c+d} \frac{c}{c+d} \right] \quad (3)$$

Because $(a + b + c + d)$ cancels, and the base rates $(a + b)$ and $(c + d)$ also cancel, we get

$$p(H|D) = a / (a + c) \quad (4)$$

Thus, with natural frequencies, the computation of the Bayesian posterior probability is reduced to the two natural frequencies a and c and the stated base rates can be ignored, unlike with conditional probabilities. Incidentally, this feature may explain why people sometimes ignore explicitly stated base rates. In the magic wand problem, a is the number of pupils with a magic hat and wand, which is four, and c is the number of students with a magic hat but without a wand, which is 12. The result is that four out of every 16 students with a magic hat can be expected to have a magic wand.

Equation 4 determines the numerical competencies required for Bayesian intuitions with natural frequencies:

- F1. The ability to understand Arabic numerals.
- F2. The ability to identify the relevant Arabic numbers a and c .
- F3. The ability to add Arabic numbers.

The competency F1 entails the ability to abstract numbers from Arabic numerals, and F2 and F3 entail reasoning processes with Arabic numbers. These three competencies are necessary for Bayesian intuitions with natural frequencies but not sufficient. In addition, natural frequencies need to be combined as according to Equation 4.

This analysis shows that natural frequencies allow for Bayesian reasoning without abilities P1 to P3. They require understanding neither probabilities nor percentages, nor their addition, multiplication, or division. F1 requires understanding natural numbers only. F2 requires ignoring the base rate information specified in the problem representation, such as five out of 20 (unlike P2), and to focus exclusively on a and c . F3 requires the ability to add natural numbers, but no multiplication or division. Division is not necessary because natural frequencies enable the posterior probability to be expressed in terms of natural numbers, that is, a out of $a + c$. Conditional probabilities, by contrast, do not allow for this feature: To report .2 out of .8 would be correct but awkward and violate the definition of fractions as the ratio of two natural numbers. In the magic wand problem, the Bayesian posterior probability corresponds to the expected number of students with a magic wand among a random sample of students with a magic hat.

Consistent with this theoretical result, the first study with adults showed that natural frequencies improved Bayesian reasoning without training, from an average of 16% (conditional probabilities) to 46% correct answers (Gigerenzer & Hoffrage, 1995). For similar experimental conditions, the meta-analysis by McDowell and Jacobs (2017) estimated an average increase from 7% to 56%.

A natural frequency representation typically begins with a prominent number as a reference class, that is, powers of 10, their halves and doubles, such as 10, 20, 50, 100, 200, 500, and 1,000. Yet using different numbers (such as 1,538 instead of 1,000) does not appear to diminish the effect of natural frequencies (Misuraca, Carmeci, Pravettoni, & Cardaci, 2009). This empirical result is consistent with the theoretical result that the size of the reference class (and the base rates as well) does not enter Equation 4.

Nevertheless, Bayesian intuitions based on natural frequencies require competencies that young children may lack. In the next section we propose a nonnumerical representation that retains the computational advantage of natural frequencies (Equation 3) but does not require understanding and adding Arabic numerals.

Bayesian Intuition and Icon Arrays

An icon array is a visual, nonnumerical representation of the result of natural sampling. It consists of *types* and *tokens*. There are

four types in an icon array, each represented by a specific icon. The tokens represent the numerosity of each type, but without Arabic numerals. To simplify counting, the individual tokens are listed next to each other. Figure 3 shows an icon array for the magic wand problem.

Icon arrays reduce the computation of the Bayesian posterior probability in the same way as shown in Equation 3. Base rates do not need to be attended to; they cancel out. In addition, icon arrays require none of the competencies P1 to P3 or F1 to F3, but only the ability to count and to know what to count. “Counting” the Bayesian posterior probability amounts to:

$$p(H|D) = \text{count}\{a - \text{tokens}\} / \text{count}\{a - \text{and } c - \text{tokens}\} \quad (5)$$

where “count{*a*-tokens}” means the count of type *a* tokens, and “count{*a*- and *c*-tokens}” means the count of tokens of types *a* and *c*. To resolve the magic wand problem, a child would need to count the students with hat and wand (four) and all students with a hat (16), and then report “four out of 16.” Icon arrays are mathematically equivalent to natural frequencies; they can be mapped onto each other one-to-one. However, icon arrays require numerical competencies developed earlier:

I1 (Counting): The ability to count tokens.

I2 (Identification): The ability to identify the correct types.

I3 (Cardinality): The ability to understand that the final count is the cardinal number of the set.

Once again, these three competencies are necessary for Bayesian intuitions with icon arrays but not sufficient. In addition, Bayesian intuitions require that the final counts are combined as according to Equation 5.

The ability to count tokens (I1) involves two principles: developing a stable counting sequence, such as one, two, three, . . . , and learning to use only one count for each token (one-to-one correspondence; Gelman & Gallistel, 1978). Stable counting sequences for small sets are usually well developed by the time children are about 2 years old, and one-to-one correspondence by about age 4 (Gelman & Gallistel, 1978). The principle of cardinality (I3) appears to be difficult only for children under 3 years of age (Sarnecka & Carey, 2008). For instance, given a picture with three birds, young children may count “one, two, three” yet still not grasp that the final count is the total number of birds. In the experiments reported in this article, the youngest children are 7 years old (second-graders); we can thus expect that I1 and I3 are part of these children’s mathematical repertoire. Because the question whether icon arrays can facilitate Bayesian intuitions in children has never been studied before, we have no evidence about children’s ability to identify the correct types in an icon array (I2).

Note that icon arrays do not require understanding how to add Arabic numbers, as with natural frequencies. A child can simply count all pupils with hats (which corresponds to *a* + *c*), rather than add the two counts of *a* and *c*. In adults, icon arrays have been shown to increase Bayesian reasoning (Brase, 2009, 2014; Cos-

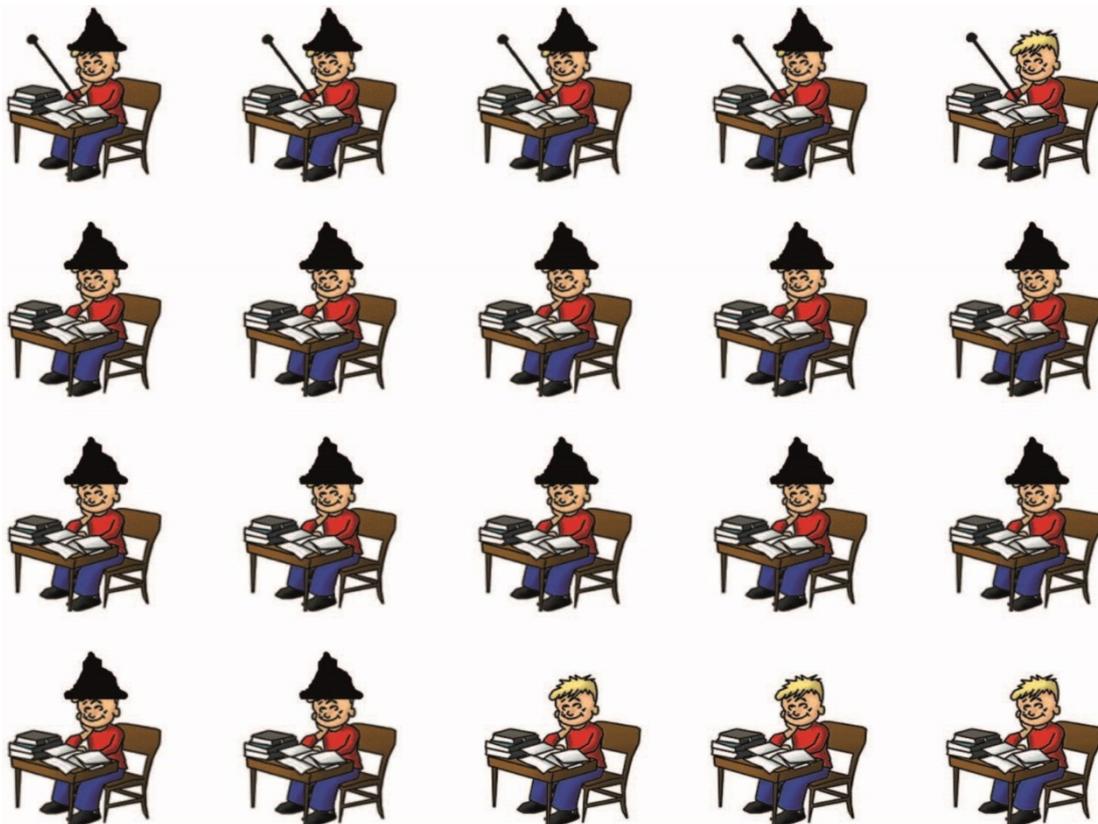


Figure 3. Icon array for the magic wand problem (from Studies 1, 2, and 3). The four types appear from top left to bottom right in the same order as the four natural frequencies in Figure 2.

mides & Tooby, 1996; Sedlmeier, 1999), and also help people with poor numeracy skills in understanding medical risk (Galesic, Garcia-Retamero, & Gigerenzer, 2009).

In summary, we distinguished three representations of numerosity: conditional probability, natural frequency, and icon array. We showed that each representation calls for a different set of numerical competencies to enable Bayesian intuitions, using Equations 1 to 5, which are mathematically but not computationally and psychologically equivalent. Solving Equation 2 requires understanding, multiplying, and adding probabilities and percentages, as defined by P1 to P3. Solving Equation 4 only requires understanding Arabic numerals and how to add them, which is a substantial facilitation. Finally, Equation 5 requires understanding neither probabilities nor Arabic numerals, but solely counting and identifying the correct types (subsets). The competencies are the preconditions for Bayesian inference, not the inference itself, which requires combining the competencies as defined in Equations 2, 4, and 5.

The key theoretical result is that external representations of numerosity are not neutral to the developing mind. Even if these are mathematically equivalent, they require different numerical abilities. Bayesian intuitions are a function of numerical competencies and representations of numerosity. This theoretical analysis allows for investigating whether children have Bayesian intuitions at an age at which they have not yet learned to reason with probabilities and Arabic numerals.

Research Questions and Overview of Studies

Studies 1 and 2 addressed the question whether icon arrays can elicit Bayesian intuitions and if so, at how early an age. We tested 7- to 9-year-olds (second- and fourth-graders), an age at which one can assume that the majority can reliably count up to the maximum numerosity used in the problems (up to 30), so that the competency I1 is satisfied. At this age, the competency I3 to understand that the final count is the cardinal number of the set is also generally available. In contrast, reasoning with Arabic numerals (as in natural frequencies) is a skill that is in the process of being developed, particularly among the younger children. As a control, we also tested children on natural frequencies without icon arrays. We did not test children with conditional probability representations, given that even older children could not solve a single problem (see Figure 1) and to spare children the frustrating experience of not being able to solve any problems given to them (Zhu & Gigerenzer, 2006). The ecological theory leads to the following predictions:

1. If children have Bayesian intuitions, then icon arrays should enable detection of them. That is, they should be able to determine the exact Bayesian posterior probability (expressed as a frequency).
2. If children have Bayesian intuitions, then icon arrays should lead to more Bayesian intuitions than would natural frequencies without icon arrays. This prediction follows from the fact that I1 to I3 develop earlier than F1 to F3.
3. The advantage of icon arrays over natural frequencies alone should be larger among the youngest group (second-graders) and smaller in the older group (fourth-graders). This prediction follows from the developmental course: To the degree that both sets of skills are available for a child, the additional facilitating effect of icon arrays should fade out.

In Study 3, we tested Predictions 1 and 2 for children with a diagnosis of developmental dyscalculia. The idea that children with dyscalculia could have Bayesian intuitions is admittedly a bold hypothesis. These children have problems with much simpler arithmetic tasks, including understanding Arabic numerals, which has been attributed in part to genetic or intrauterine causes. Yet the ecological theory shows that counting abilities suffice for Bayesian intuitions when using icon arrays and that reasoning with Arabic numerals is not necessary. That is, if dyscalculia is caused in part by difficulties in reasoning with Arabic numerals, icon arrays should reduce the differences between children with and without dyscalculia. A second motivation for this hypothesis is that for low-numerate adults, natural frequencies reduce the gap in probabilistic reasoning between them and high-numerate adults (Galesic et al., 2009; Johnson & Tubau, 2013). The prediction is:

4. If children diagnosed with dyscalculia have Bayesian intuitions, then icon arrays should enable detection of them. These children should be able to identify the correct types (I2), and combine them according to Equation 5. In addition, icon arrays should lead to larger proportions of Bayesian intuitions than do natural frequencies without icon arrays.

In the final study, we investigated the development of Bayesian intuitions in older children in Grades 5 to 7, with natural frequencies alone. At this age, the preconditions F1 to F3 should be in place for the majority of children. The prediction is:

5. From fifth- to seventh-graders, an increasing proportion of children can find the exact Bayesian answer without icon arrays, using natural frequencies alone.

General Methodological Principles

All studies have in common the following:

1. None of the children had received education in mathematical probability or Bayes' rule in school.
2. All problems used numerosities ≤ 30 for second- and fourth-graders and ≤ 100 for fifth- to seventh-graders.
3. Children were asked for the posterior probability in the format " x out of y " because they did not know how to calculate ratios or probabilities between 0 and 1. Children with dyscalculia received the same response format, but in addition a simpler one that solely required identifying the relevant tokens by marking these with a pen rather than counting (Study 3).
4. A child's response was classified as a Bayesian intuition only if it was the numerically exact Bayesian response. For instance, in the magic wand problem, only the exact response "four out of 16" was classified as a Bayesian

intuition. In contrast, slight deviations such as “five out of 16” were not. The reason is a conservative one, namely to avoid misclassifying a non-Bayesian intuition as a Bayesian one. In classifying children’s judgments, one can make two errors: a false positive, that is, to classify a non-Bayesian intuition as Bayesian, or a miss, that is, to classify a Bayesian intuition as non-Bayesian. The strict classification criterion used here minimizes false positives although it may miss some Bayesian intuitions distorted by calculation error (McDowell, Galesic, & Gigerenzer, 2018). Because we test hypotheses that are new, it is only reasonable to score conservatively against Predictions 1 to 5. To check whether this conservative criterion overlooks mere counting errors, we also provide a separate analysis of counting errors.

5. The Ethics Committee of the Max Planck Institute for Human Development (Germany) approved the methodology.

Study 1: Do Second- and Fourth-Graders Have Bayesian Intuitions?

Method

In this study, we tested Predictions 1 to 3 for second- and fourth-graders. This is the first study to have tested Bayesian intuitions at this young age and also the first to have employed icon arrays with children. To determine the effect of icon arrays beyond natural frequencies alone, children were randomly assigned to two conditions: natural frequencies or natural frequencies plus icon arrays. Because quantitative Bayesian intuitions may prove difficult for children at this young age, we also included a simpler test of qualitative judgments. In a qualitative test, children are asked whether a hypothesis is more likely than another, given the data, rather than being asked for a quantitative estimate. We cautiously refer to these as *qualitative judgments* instead of qualitative Bayesian intuitions because, unlike (quantitative) Bayesian intuitions that require an exact number, the correct direction (e.g., more or fewer students with a magic wand among those with a magic hat) can be guessed with a probability of .5 and may result from non-Bayesian rules (McDowell et al., 2018). We asked the qualitative questions in a single-event and a frequency format. For the magic wand problem, the two qualitative questions were:

Single-event question. If someone in that school wears a magic hat, do you think he also has a magic wand?

- a. Likely yes.
- b. Likely no.

Frequency question. Imagine a group of students from that school who are all wearing a magic hat. Are there more students with a magic wand or more students without a magic wand?

- a. More students with a magic wand.
- b. More students without a magic wand.

The motivation for including these two variants is an issue raised by Piaget and Inhelder (1951/1975) at the very end of their book: Is probability about single events or about frequencies?

Piaget and Inhelder located this question at the center of the theory of applied probability, which divided the statisticians Richard von Mises and Hans Reichenbach from Emile Borel (and, we might add here, the subjective Bayesian school). Piaget and Inhelder held that people reason about single events by relating these to a class of events, following Reichenbach’s version of the frequency view. At the same time they proposed that a psychological analysis of children’s probabilistic intuitions “can help us solve this question” (p. 243). We do not believe, however, that analyzing children’s intuitions can answer this centuries-old philosophical question that continues to be debated to the present day.

What the ecological analysis instead suggests is a hypothesis for whether children respond differently to these two formats. Finding the answer to the frequency question requires two counts to determine which is higher. Finding the answer to the single-event question requires one further step to realize that this difference in counts can be also applied to a single person. Thus, to the degree the solution is based on counting rather than reasoning with Arabic numerals, the frequency question should elicit more responses consistent with Bayes’ rule. If this is correct, we expect that (a) the frequency question results in a larger proportion of Bayesian intuitions than the single-event question, and (b) this difference is larger for younger children than for older ones.

Participants. Ninety-one second-graders (45 girls and 46 boys) and 85 fourth-graders (45 girls and 40 boys) were recruited from public elementary schools in Berlin. The mean age of the second-graders was 7;6 years, range 7;1 to 10;0 years, and the mean age of the fourth-graders was 9;6 years, range 9;0 to 11;1 years.

Materials and procedure. Children were tested in their classrooms in small groups of two to six. We constructed six Bayesian reasoning problems with content suitable for young children. All problems were tested in pilot research to be of interest to children at this age. One was the magic wand problem; the others are shown in the Appendix. Every child received a booklet, with each problem on a separate page. The cover page showed the Sesame Street character Count von Count surrounded by numbers together with the text: “Count von Count needs your help! He is supposed to urgently answer important questions, but just now he mixed up his numbers and is confused. Can you help him?” To explain the unfamiliar task, the experimenter read aloud an example problem (not included in the six test problems), the correct answer, and a short explanation. Then the children were instructed: “Now you have to solve similar problems on your own. Read each problem carefully and answer all three questions. If you do not know the answer, pick what is most plausible to you.”

After the children read a problem, they were asked three questions. The first two were the qualitative questions, the third the quantitative question about the posterior probability. Children were invited to raise their hands if they did not understand the instruction, but this rarely happened even with second-graders. If so, the experimenter went to the child and quietly and slowly read the question to the child without rephrasing. The teacher was present in the room but did not intervene. Children were randomly assigned to two conditions in a between-subjects design: natural frequencies and icon arrays. The natural frequency condition provided text but no icon arrays. The icon array condition always showed the icon array together with the natural frequency repre-

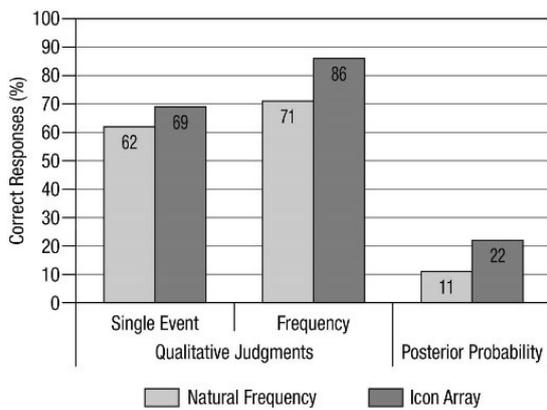


Figure 4. Percentage of Bayesian intuitions (posterior probability) and qualitative judgments among children in second grade, for natural frequencies and icon arrays.

An individual analysis showed that with icon arrays, seven second-graders (16%) showed Bayesian intuitions for most of the problems, that is, for more than three problems. With natural frequencies, this was observed for two second-graders (4%).

Qualitative judgments. In all four conditions, natural frequencies versus icon arrays and single event versus frequency questions, qualitative judgments were above chance level (all p s < .001, chance level = 50%; Figure 4). More correct responses were observed when the question was directed at a group (frequency) rather than a person (single event) in both the natural frequency ($W = 1290, p = .006$) and the icon array condition ($W = 1284, p < .001$). Icon arrays improved correct answers for single events (icon array: median = 1, IQR = 0–2; natural frequency: median = 0, IQR = 0–1, $W = 1317.5, p = .014$) and frequency judgments (icon array: median = 6, IQR = 5–6, $W = 1097, p = .036$).

Fourth-graders. Forty fourth-graders participated in the natural frequency condition and 45 in the icon array condition.

Bayesian intuitions. The fourth-graders gave the exact Bayesian response more frequently in the icon array condition (median = 4, IQR = 3–5) than in the natural frequency condition (median = 2.5, IQR = 0.75–4, $W = 1191, p = .001$). An individual analysis showed that with icon arrays, 30 fourth-graders (67%) found the exact Bayesian posterior probability for more than three problems. For natural frequencies, this number was 15 (38%).

Qualitative judgments. In all four conditions, natural frequencies versus icon arrays and single event versus frequency questions, qualitative judgments were above chance level (all p s < .001). The median number of correct qualitative judgments ranged between 5.25 and 6 (see Figure 5). The amplifying effect of icon arrays and frequency questions is no longer visible, consistent with the hypothesis that the advantage of frequency questions over single event questions fades with increasing age.

Discussion

Study 1 demonstrates for the first time quantitative Bayesian intuitions among second-graders. In the icon array condition, children found the exact Bayesian posterior probability for 22% of all problems. This number increased to 61% among fourth-graders. Consistent with Prediction 1, icon arrays could elicit Bayesian intuitions in children who had had no teaching in probability theory or Bayes' rule. In both age groups, icon arrays also resulted in larger proportions than did natural frequencies, consistent with Prediction 2. The results for Prediction 3 are not as clear-cut but show the predicted tendency. Among the second-graders, the ratio of Bayesian intuitions in the two conditions is 2 to 1 in favor of icon arrays, whereas among fourth-graders it is 1.4 to 1. That is, the advantage of icon arrays, as measured by the ratio of Bayesian intuitions, appears to diminish slightly among the older group.

Study 1 also shows for the first time qualitative judgments in both second- and fourth-graders. Icon arrays increased qualitative judgments in second-graders, whereas this additional effect was no longer visible in fourth-graders, who reached a level of around 90% correct answers.

The additional hypothesis investigated concerned the phrasing of the qualitative question. The difference mattered only for the younger group, where frequency questions elicited on average more Bayesian intuitions than single-event questions, 78% and 65.5%, respectively, across all problems and representations (see Figure 4). No such difference could be observed among the fourth-graders. As mentioned above, Piaget and Inhelder suggested that psychological tests might settle the controversy about whether probability applies to singular events. We do not think so. Rather, the results suggest that frequency questions facilitate qualitative judgments among the 7-year-olds, but that this effect disappears by the age of 9, consistent with the ecological analysis.

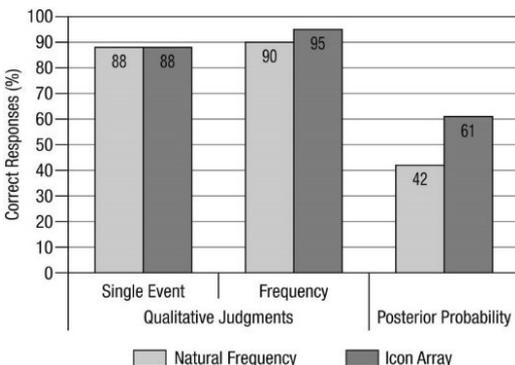


Figure 5. Percentage of Bayesian intuitions (posterior probability) and qualitative judgments among children in fourth grade, for natural frequencies and icon arrays.

sentation, which allowed the problem and the question to be formulated in a standardized way. Frequencies of correct responses were tested using Wilcoxon's rank sum tests because scores were not normally distributed within grades and conditions as assessed by Kolmogorov–Smirnov tests. For the magic wand problem, the quantitative question is provided above; the questions for all other problems are provided in the Appendix.

Results

Second-graders. Forty-eight second-graders participated in the natural frequency condition and 43 in the icon array condition. Each child worked on six problems.

Bayesian intuitions. The second-graders found the exact Bayesian posterior probability more frequently for problems in the icon array condition than in the natural frequency condition (icon array; median = 1, IQR = 0–2; natural frequency: median = 0, IQR = 0–1, $W = 1317.5, p = .014$; Figure 4).

Study 2: Replication of Study 1

Method

In Study 1, we showed for the first time that Bayesian intuitions can be elicited among second-graders, consistent with the analysis of numerical competencies for icon arrays. Yet this result appears to conflict with the literature that found even most adults failing to solve Bayesian problems (e.g., Kahneman & Tversky, 1972; Kahneman, 2011). The children in Study 1 might have been extraordinarily gifted with probabilistic intuitions, or we might have seen an inexplicably large number of false positives, that is, children who found the Bayesian solution by chance (for an analysis of this probability, see the Discussion section below). Study 2 therefore tested whether the results could be replicated with a different group of children.

Participants. Thirty-four second-graders (17 girls and 17 boys) and 32 fourth-graders (14 girls and 18 boys) from a public elementary school in Berlin participated in Study 2. The mean age of the second-graders was 7;6 years (range 6;10 to 9;1), and the mean age of the fourth-graders was 9;8 years (range 8;1 to 10;10 years). None of the children had participated in Study 1.

Materials and procedure. Materials and procedure were identical to those in Study 2.¹

Results

Figures 6 and 7 show that Bayesian intuitions could be elicited in the new group of children and in both age groups, replicating Study 1.

Bayesian intuitions. The second-graders gave the exact Bayesian response more frequently in the icon array condition than in the natural frequency condition (icon array: median = 2, IQR = 1–2.5; natural frequency: median = 0, IQR = 0–9, $W = 258, p < .001$), consistent with Predictions 1 and 2 (see Figure 6). Among fourth-graders, the difference increased (icon array: median = 3, IQR = 3–4; natural frequency: median = 0, IQR = 0–1, $W = 229, p < .001$, see Figure 7). Prediction 3 says that the advantage of icon arrays over natural frequencies diminishes with age. Consistent with this prediction, the advantage is about 10:1 for second-graders and 4:1 for fourth-graders, as measured by the ratio of the total number of Bayesian intuitions with icon arrays to those with natural frequencies. All in all, the Bayesian intuitions among second-graders were more frequent than in Study 1 in the icon array condition, whereas those of the fourth-graders were slightly lower, but still above 50%.

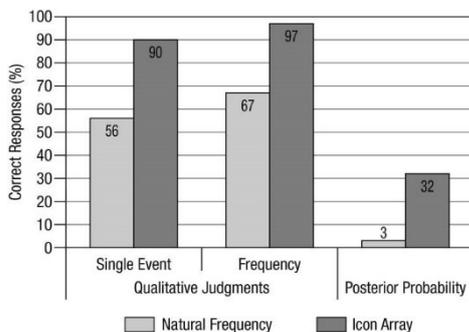


Figure 6. Percentage of Bayesian intuitions (posterior probability) and qualitative judgments among children in second grade, for natural frequencies and icon arrays (Study 2).

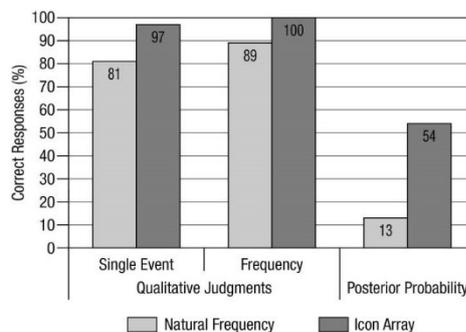


Figure 7. Percentage of Bayesian intuitions (posterior probability) and qualitative judgments among children in fourth grade, for natural frequencies and icon arrays (Study 2).

Qualitative judgments. Among second-graders, the average proportion of qualitative judgments across all four conditions was in the same range as in Study 1, whereas the effect of icon arrays was larger in Study 2, on average about 15 percentage points. Similarly, the positive effect of frequency questions replicates, as does its relative decline with age. As in Study 2, qualitative judgments are close to perfect among the fourth-graders, in particular with icon arrays.

Individual differences. Study 2 also replicates the existence of substantial individual differences within both age groups. Tables 1 and 2 show these differences, summarized across Studies 1 and 2. Individual differences occur with both icon arrays and natural frequencies. With icon arrays, the majority of the second-graders (71%, 44 out of 62) could solve one or more problems (Table 1, “Correct response”). Eighteen out of 62 (29%) could not solve a single problem, while 18 (29%) solved one problem, 12 (19%) solved two problems, four (6%) solved three problems, and 10 (16%) solved four problems. For comparison, with natural frequencies alone, the majority of second-graders (67%; 42 out of 63) could not solve a single problem in the natural frequency condition, 11 solved one, eight solved two, and one each solved four and five problems.

With icon arrays, the distribution of the number of exact Bayesian responses among fourth-graders (Table 2, “Correct response”) centers on three to five correct answers, with a mode of four (out of six problems). There were even some fourth-graders who solved the entire set of problems. All in all, 59% of the fourth-graders solved more than half of the problems, and the same percentage of all problems were solved. For comparison, with natural frequencies, the distribution of Bayesian intuitions is left-skewed with a mode of 0. Altogether, 29% of fourth-graders solved more than half of the problems, and 34% of the problems were solved.

It might be of interest to know how many students gave answers close to the correct response. Specifically, a student might identify

¹ We also used the replication to check whether the formulation “Think of a group of students . . .” compared with “Think of all students . . .” made any difference, but these had no effect on children’s Bayesian intuitions.

Table 1
Individual Differences in Bayesian Intuitions Among Second-Graders (Studies 1 and 2 Combined)

Condition	0	1	2	3	4	5	6	Bayes (%)
Icon array ($n = 62$)								
Correct response	18	18	12	4	10	0	0	94 (25%)
Tolerance of ± 1 error	10	20	15	5	7	5	0	118 (32%)
Natural frequency ($n = 63$)								
Correct response	42	11	8	0	1	1	0	36 (10%)
Tolerance of ± 1 error	24	22	8	7	1	1	0	68 (18%)

Note. Shown is the number of second-graders who correctly solved 0, 1, . . . , 6 Bayesian problems in the icon array and the natural frequency condition. For instance, in the icon array condition, 18 children solved none of the six problems, 18 solved one problem, and 12 solved two problems. The Tolerance columns show the same analysis except that a counting (or adding) error of plus or minus one is tolerated (see text). The last column reports the number (percentage) of all problems for which children had Bayesian intuitions.

the correct subsets in the icon array but make a counting error. A counting error is most likely to occur when counting the larger numerosity in “ x out of y ,” resulting in “ x out of $y - 1$ ” or “ x out of $y + 1$.” In the magic wand problem, this would result in the responses “4 out of 15” or “4 out of 17.” We determined the equivalent responses for all six problems, and then analyzed the number of children who actually gave these responses. Tables 1 and 2 (“tolerance”) show the result. In the icon array condition, Bayesian responses increased by seven and five percentage points for second- and fourth-graders, respectively. For control, we made the same analysis in the natural frequency conditions, where adding, not counting is a possible source of error. Here, the increase was eight percentage points for both groups. Thus, allowing for counting (adding) errors would lead to a moderately higher estimate of Bayesian intuitions. Yet we cannot know whether this increase is entirely due to errors in counting or adding, meaning that these figures may provide an upper estimate of Bayesian intuitions distorted by error.

Like the analysis of means, the analysis of individual differences supports Predictions 1 and 2. It also suggests that the advantage of icon arrays over natural frequencies diminishes with age, from a ratio of 2.5 to one of 1.7, consistent with Prediction 3 (Tables 1 and 2). A close inspection of the differences between the two age groups shows that for icon arrays, the mode of the distribution shifts from the left to the middle, where it centers on four correct responses. In the natural frequency condition, the distribution similarly moves to the right from second-graders to fourth-graders, but with a substantial delay.

Discussion

Could the response pattern in Tables 1 and 2 be the product of mere guessing by picking numbers randomly rather than instances of Bayesian intuitions? To answer this question, we need to ask whether there are more correct responses than would be expected if the children had guessed. Unlike for qualitative judgments, where chance is 50%, it is not easy to define the result of guessing for quantitative Bayesian intuitions. A first definition would be that children randomly pick two numbers x and y to answer the question “___ out of ___.” In this definition, there are infinite possibilities and it appears virtually impossible for a child to guess the exact Bayesian posterior probability. Yet it is reasonable to constrain the numbers x and y in two ways. First, we assume $x < y$, a regularity that children consistently followed; second, we assume $x < y \leq 30$, which is the largest numerosity in the problems used. Given these two constraints, there are 435 possibilities for picking numbers by chance.² Assuming a uniform distribution and given the strict classification of a Bayesian intuition in terms of an exact number, we can expect one random hit in 435 problems.³ We can now compare this expectation with the actual performance of the second-graders. In Studies 1 and 2, a total of 125 second-graders worked on six problems each, resulting in 750 problems. This leads to an expected number of one to two correct solutions by chance; the actual number of correct solutions was 130 (see Table 1). Fourth-graders worked on 702 problems, which also leads to an expected number of one to two correct solutions by chance; here, the actual number of correct solutions was 329.

A second way of formulating a chance hypothesis is to assume that children use only the five numbers provided in each problem (see the natural frequency version of the magic wand problem) and combine them in a random way. As a conservative constraint we use the observation that children typically used no more than one or two of the five numbers to determine x or y (if we allowed for three of the five numbers, which occurred in rare cases [see below], the possibilities of random choice would be even higher). If children used two numbers, they virtually always added them up (see the section on non-Bayesian rules below). For x and y , this results in five single numbers plus $5 \times 4/2 = 10$ pairs (order does not matter because addition is commutative), that is, a total of 15 possible numbers each. With the constraint $x < y$, there are $15 \times 14/2 = 105$ possibilities. Assuming a uniform distribution, we can then expect one random hit in 105 problems. Given that the second-graders worked on 750 problems, this chance hypothesis would expect seven to eight correct solutions; as mentioned above, the actual number of correct solutions was 130. For fourth-graders, the chance hypothesis would expect six to seven correct solutions, in face of 329 actual solutions. Hence, neither of these two chance hypotheses can explain the large number of observed Bayesian intuitions in second- and fourth-graders.

² For natural numbers 1 to N , there are $N(N - 1)/2$ ordered pairs that satisfy the constraint that the first number is smaller than the second. For $N = 30$, this results in $30 \times 29/2 = 435$ pairs.

³ We assume here that children do not transform the correct answer into an equivalent ratio by multiplication or division, such as four out of 16 into one out of four, given that this is beyond the arithmetic capacities of young school children and was not observed in the actual estimates.

Table 2
Individual Differences in Bayesian Intuitions Among Fourth-Graders (Studies 1 and 2 Combined)

Condition	0	1	2	3	4	5	6	Bayes (%)
Icon array (<i>n</i> = 61)								
Correct response	3	3	7	12	20	13	3	216 (59%)
Tolerance of ± 1 error	1	3	4	12	23	14	4	233 (64%)
Natural frequency (<i>n</i> = 56)								
Correct response	19	10	6	5	8	4	4	113 (34%)
Tolerance of ± 1 error	7	17	6	6	11	5	4	140 (42%)

Note. Shown is the number of fourth-graders who correctly solved 0, 1, . . . , 6 Bayesian problems in the icon array and the natural frequency condition. For instance, in the icon array condition, three children solved zero problems and 20 solved four problems. The Tolerance columns show the same analysis except that a counting (or adding) error of plus or minus one is tolerated. The last column reports the number (percentage) of all problems for which the children had Bayesian intuitions.

In sum, Study 2 replicates the major finding of Study 1 that icon arrays can elicit Bayesian intuitions already in second- and fourth-graders. Among second-graders, the percentage of Bayesian intuitions even reached 32% in the icon array condition, albeit only 3% in the natural frequency condition.

Study 3: Do Children With Dyscalculia Have Bayesian Intuitions?

The hypothesis that children with dyscalculia could have Bayesian intuitions may appear to have a low prior probability. As a first approximation, dyscalculia is the equivalent of dyslexia in the domain of numbers. It is often suspected to have a genetic component and linked to a reduction of gray matter density in the left intraparietal sulcus, the location where brain activity is observed during mental arithmetic (Dehaene, 1977/2011). The fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*; American Psychiatric Association, 2013) estimates that 5% to 15% of school-age children may suffer from a specific learning disorder that hampers numerical competency. Developmental dyscalculia is one of these disorders, estimated to occur in about 3% to 6% of the general population (Skagerlund & Träff, 2016). Although its consequences appear to be about as detrimental as those of dyslexia, there is lower awareness of it in the general population and comparatively less research (Beddington et al., 2008).

Dyscalculia manifests itself as a difficulty in learning and comprehending arithmetic and in executing calculation procedures (Butterworth, 2005; Dehaene, 1997/2011). Two subtypes based on their etiology have been distinguished (von Aster & Shalev, 2007). First, *pure developmental dyscalculia* is assumed to result from genetic or intrauterine adverse events that lower the basic understanding of quantity. These children have problems with both symbolic and analog quantities. This form of dyscalculia is associated with genetic anomalies such as Turner's syndrome and phenylketonuria and with intrauterine events such as alcohol exposure; in addition, 11% of children with dyscalculia are reported to have attention-deficit/hyperactivity disorder (ADHD; Soares & Patel, 2015). Moreover, cerebral lesions selectively impair mental calculation, although these are quite infrequent (Dehaene, 1997/2011, Chapter 7). The second subtype, *comorbid dyscalculia*, assumes a deficit in symbolic number processing, which may be the consequence of comorbid dyslexia (von Aster & Shalev, 2007). In contrast to the first group, children with comorbid dyscalculia appear to have an analog understanding of numerosity but cannot connect this analog understanding to a number symbol.

In addition to the uncertainty about classification, the mechanisms underlying developmental dyscalculia are not well understood. Skagerlund and Träff (2016) distinguish between two hypotheses. One is based on the research postulating an approximate number system (ANS) that is responsible for representing large and approximate numbers by an analog number line that is roughly logarithmic in nature (Dehaene, 1997/2011). The hypothesis is that developmental dyscalculia is a consequence of a defective ANS. In contrast, the *access deficit hypothesis* states that the problem is not with processing numerosities in general but with accessing magnitude information from symbols, that is, numerals (Rousselle & Noël, 2007). This hypothesis suggests that children with developmental dyscalculia have problems mainly with symbolic magnitudes (numerals) but are capable of performing nonnumerical tasks because the latter does not require access to the representation of magnitude underlying Arabic numerals (Skagerlund & Träff, 2016).

Tests and studies of dyscalculia have focused on various aspects, including basic number knowledge, simple one-digit arithmetic, multidigit calculation, and arithmetic fact retrieval (Skagerlund & Träff, 2016). To the best of our knowledge, no study has investigated Bayesian intuitions or applied the present ecological perspective. If dyscalculia were attributable to a genetic anomaly or brain defects caused by adverse intrauterine events, then such a study should show consistently negative effects.

Method

The hypothesis that children with dyscalculia also have Bayesian intuitions can be derived in the following way. To the degree that the cause of dyscalculia lies in a difficulty with accessing magnitude information from symbols such as Arabic numerals rather than with processing numerosity in general, it should be possible to elicit Bayesian intuitions with icon arrays, at least in some children (Prediction 4). This prediction is independently motivated by observations with low-numerate adults. Their performance was much closer to that of high-numeracy adults when given natural frequencies as opposed to probabilities (Galesic et al., 2009; Galesic, Gigerenzer, & Straubinger, 2009; Johnson &

Tubau, 2013). Thus, just as natural frequencies appear to even out individual differences between adults varying in numeracy, icon arrays could do the same between children with and without dyscalculia.

To test Prediction 4, we took care to adapt the problems to the numerical abilities of children diagnosed with dyscalculia, using a modified icon array condition and natural frequency condition. In both conditions, we replaced the qualitative single-event question with a nonnumerical analog task, which does not require counting but simply marking tokens with color pens. This task measures two elementary preconditions for Bayesian intuitions with icon arrays: to identify the correct tokens given a type and to translate natural frequencies into tokens and types.

Modified icon array condition. In the icon array condition, the magic wand problem (along with all other problems) was adapted by eliminating all Arabic numerals in the text, leaving only the indeterminate quantity pronoun *some*:

Of the students at Hogwarts School of Witchcraft, some have a magic wand. Of these, some also have a magic hat. Also, some of the students who do not have a magic wand have a magic hat. Here you see some of the children from the school.

At this point, the icon array (see Figure 3) was shown and the qualitative frequency question was asked, as in Studies 1 and 2. This was followed by the new analog task:

Now take a green pencil and circle all students with a magic hat. Then take a red pencil and cross out every student who wears a magic hat and has a magic wand.

The analog task also contains no Arabic numerals. It does not directly measure Bayesian intuitions, but rather two preconditions: the ability to identify the types and tokens necessary for Equation 5. Finally, the quantitative question was asked, identical to the one asked in Studies 1 and 2. The quantitative question should be the more difficult one, whereas the analog task should be relatively easier.

Modified natural frequency condition. The natural frequency condition was similarly modified to include an analog task. Unlike the icon array condition, it provided natural frequency information, but in a simplified format:

Out of every 20 students at Hogwarts School of Witchcraft, four have a magic wand and a magic hat, one has a magic wand but no magic hat, 12 do not have a magic wand but have a magic hat, and three do not a magic wand and also do not have a magic hat.

In this simplified “short menu,” only the four natural frequencies are presented, not the base rates (Gigerenzer & Hoffrage, 1995). In adults, the short menu elicited more correct Bayesian responses than the one used in Studies 1 and 2 (which directly corresponds to the conditional probability format). The modified condition still requires processing of Arabic numerals.

The first question was the qualitative frequency question, as in Studies 1 and 2. This was followed by an analog task that contained “empty icons” (without hats and wands, that is, no types) rather than an icon array (see Figure 8). Participants were told: “Here you see 20 of the students. But who has a magic wand, and who a magic hat? Now take a green pencil and circle all students



Figure 8. Analog task containing “empty icons” for Study 3.

with a magic hat. Then take a red pencil and cross out every student who wears a magic hat and has a magic wand.”

As in the icon array condition, the analog task reflected the special difficulties of children with dyscalculia by allowing them to provide an answer without using number symbols. The final question was the quantitative estimate of the Bayesian posterior probability, as in Studies 1 and 2 and in the icon array condition of Study 3.

In sum, the icon array condition was modified so that no Arabic numerals were presented and an additional analog task was introduced that tests the basic ability to identify the correct types and tokens in the array. The natural frequency condition was modified so that natural frequencies were presented in the “short menu,” and an analog task was introduced that measures the ability to identify the tokens necessary for calculating Equation 5. This requires processing of Arabic numerals and thus can be expected to be more demanding for children with dyscalculia.

Participants. Thirty-seven children (29 girls and 8 boys) from second to sixth grade with a diagnosis of dyscalculia according to *DSM-IV* participated in Study 3. The children were recruited through their schools or therapists with the written informed consent of their parents. As in the previous studies, each child was randomly assigned to either the icon array or natural frequency condition. There were 20 children in the icon array condition with an average age of 9;10 years ($SD = 1;7$), 17 female and three male, and a mean IQ of 99 ($SD = 13$). One of the children also had a diagnosis of ADHD. In the natural frequency condition, there were 17 children with an average age of 9;7 years ($SD = 1;7$), 12 female and five male, and a mean IQ of 95 ($SD = 10$). One of the children was also diagnosed with dyslexia.

Materials and procedure. All children were tested individually in a quiet room for approximately 45 min. As in Studies 1 and 2, each child was first read an example problem and then worked on the same six Bayesian problems. Then three questions were asked: the qualitative frequency question, followed by the analog task and then the quantitative question. After the children had worked on the six problems, they were given the short version of the Culture Fair Intelligence Test (CFT 20-R; Weiss, 2006), which is designed for people with language comprehension difficulties and assesses formal-logical operations along with the ability to recognize and transform spatial configurations. Finally, they were given the subtest “quantity estimation” of the arithmetic abilities test RZD 2–6 (Jacobs & Petermann, 2005), which assesses the acuity of analog quantity representations and allowed for a check of the children’s diagnosis of dyscalculia. The performance of the two experimental groups (icon array vs. natural frequency) did not differ in these tests. Children received 10 euros for expenses.

Results

Bayesian intuitions. With icon arrays, children with dyscalculia succeeded in finding the exact Bayesian posterior probability for 50% of the problems (see Figure 9). This performance approaches the level of children without dyscalculia at the same average age, that is, fourth-graders (59%, Table 2). With natural frequencies, children could find the exact Bayesian response for 20% of the problems, which is lower than the average among fourth-graders without dyscalculia (34%, Table 2).

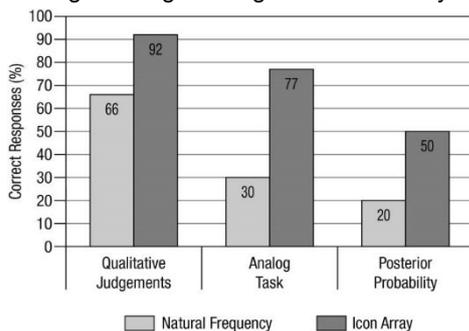


Figure 9. Bayesian intuitions can be elicited in children diagnosed with dyscalculia (*DSM-IV*). The icon array condition and the natural frequency condition have been modified to fit the numerical competency of the children (see text). The qualitative judgments are identical to the frequency question in Studies 1 and 2, and the posterior probability question is also identical. The analog task is specifically adapted to children with dyscalculia (see text).

Analog task. The analog task does not measure Bayesian intuitions directly but instead two preconditions. In the modified icon array condition, where an icon array was presented, 77% were able to identify the correct tokens, that is, all tokens necessary for the Bayesian response. In the modified natural frequency conditions, where no icon array was shown (only “empty icons”), only 30% were able to translate the Arabic numerals into types and tokens in an empty array. Both tasks measure the ability to identify the types and tokens whose relation determines the Bayesian posterior probability.

Qualitative judgments. With icons and without Arabic numerals, the qualitative judgments were correct 92% percent of the time, which is close to the average performance of fourth-graders. With natural frequencies, the performance of the children was slightly above chance (50%), indicating again a difficulty in understanding Arabic numerals. The natural frequency condition is the only one where the expected poor performance of children diagnosed with dyscalculia could be consistently found.

Association between Bayesian intuitions and IQ. We assessed IQ using the Culture Fair Intelligence Test (CFT 20-R). The mean was 97.4 (median 97), with a range from 67 to 124 points. The Spearman rank correlation between IQ and exact quantitative Bayesian intuitions was .57 ($p = .04$) in the natural frequency group and .36 ($p = .16$) in the icon array group. This result, limited by the small sample size, shows a similar tendency to that observed with the children in Studies 1 and 2. Icon arrays tend to equalize individual differences in Bayesian intuitions between children, whereas natural frequencies make these more pronounced. This result is consistent with the theoretical analysis showing that icon arrays require numerical competencies that are available earlier in development.

Individual differences. Could the substantial proportion of Bayesian intuitions among children diagnosed with dyscalculia be attributable to a few outliers? Table 3 (“correct responses”) shows that this is not the case, similar to children without dyscalculia. With icon arrays, there were three children who found the correct response for all six problems but an equal number of children who

Table 3
Individual Differences in Bayesian Intuitions Among Children Diagnosed With Dyscalculia

Condition	0	1	2	3	4	5	6	Bayes (%)
Icon array (<i>n</i> = 20)								
Correct response	1	5	3	3	3	2	3	60 (50%)
Tolerance of ± 1 error	0	3	5	2	2	5	3	70 (58%)
Natural frequency (<i>n</i> = 17)								
Correct response	5	7	2	3	0	0	0	20 (20%)
Tolerance of ± 1 error	2	9	3	2	1	0	0	25 (25%)

Note. Shown is the number of children who correctly solved 0, 1, . . . , 6 Bayesian problems in the icon array and the natural frequency condition. For instance, in the icon array condition, one child solved none of the six problems, and three all six problems. The Tolerance columns show the same analysis except for that a counting (or adding) error of plus or minus one is tolerated. The last column reports the number (percentage) of all problems for which children had Bayesian intuitions.

answered two, three, or four of the problems correctly. If we tolerate an error (see Tables 1 and 2), the proportion of Bayesian intuition would increase slightly from 50% to 58% with icon arrays, and to 25% for the natural frequency condition.

Discussion

Teachers and therapists had expected that children diagnosed with dyscalculia would not be able to solve a single Bayesian problem. The surprising result is that with icon arrays, these children exhibited Bayesian intuitions for 50% of a total of 120 problems. This performance approximates the level of children without dyscalculia of the same average age, that is, fourth-graders.

Equally interesting are the tasks at which these children failed. The major failures are in qualitative judgments and in translating numerals into tokens in an empty icon array. Both occur in the same situation: when Arabic numerals are presented. In combination with the rather impressive performance with icon arrays, this suggests that part of the impairment is in processing and reasoning with Arabic numerals, such as natural frequencies.

These results cannot by themselves disclose the nature of dyscalculia yet may reveal a major factor. If dyscalculia were attributable to a genetic or intrauterine event, as suggested by pure developmental dyscalculia, one should not obtain the present result that icon arrays lead to almost similar Bayesian intuitions in children with and without dyscalculia. Instead, the difference between the surprising ability to identify the exact posterior probability based on nonnumerical information and the equally surprising failure to reason with natural frequencies suggests that part of the difficulty experienced by these children is in dealing with Arabic numerals. If this is correct, then children with dyscalculia lack the competencies F1 to F3, but not I1 to I3.

Although some explanations imply that children with dyscalculia have severe difficulties in translating analog *internal* representations of numbers into symbolic ones (von Aster & Shalev, 2007), the ecological analysis suggests the reverse hypothesis. The problem may instead be located in difficulties in translating *external* representations of numerosity into internal ones. This impairment corresponds to a lack in what Gelman and Gallistel (1978) call *number abstraction*. In contrast, Bayesian intuitions emerge if the external representation uses types and tokens rather than Arabic numerals.

There is an important caveat to this conclusion. Developmental dyscalculia is a heterogeneous condition that is not as easily measurable as limited eyesight, for instance, and we had to rely on existing diagnoses according to *DSM-IV*. Therefore, we checked these diagnoses by testing each child on a subtest of the arithmetic abilities test for Grades 2 to 6 (RZD 2–6, Jacobs & Petermann, 2005), which assesses the ability for analog quantity representations. If one classifies children whose performance is two standard deviations below their respective grade average as impaired, then this diagnosis is consistent with the *DSM-IV* diagnosis in only about half of all cases. Despite that difference, however, a reanalysis showed that those classified by the RZD criterion were indistinguishable from the other half in terms of Bayesian intuitions across icon array and natural frequency tasks. Thus, the conclusions are robust with respect to this diagnostic check.

Study 4: The Development of Bayesian Intuitions in Fifth- to Seventh-Graders

Method

In the final study, we investigated the development of Bayesian intuitions in older children without training in probability theory, up to seventh grade. Here, the focus was on Bayesian intuitions elicited by natural frequencies alone, that is, without icon arrays. At this age, the numerical competencies F1 to F3 should be in most children's mathematical repertoire. The purpose was to test Prediction 5 that in this age group an increasing number of children can find the exact Bayesian posterior probability (expressed as frequency) with natural frequencies alone.

There exists a single study that showed substantial Bayesian intuitions in children (in China) when given information in natural frequencies (Zhu & Gigerenzer, 2006), which could not be replicated with Italian children (Pighin, Girotto, & Tentori, 2017). This difference was tentatively attributed to the gap in mathematical achievements between East Asian and Western children. East Asian children have indeed repeatedly outperformed Western children, including Italian and German children, in mathematics (Artelt, Demmrich, & Baumert, 2001). Moreover, their linguistic number systems are easier and more logical than Western ones. However, if the above theoretical analysis is correct, and the assumptions F1 to F3 are in place, mathematical achievement beyond these competencies should make little difference for

Bayesian intuitions. Thus, Prediction 5 should apply equally for Western (here: German) children and Chinese children. Moreover, the effect size should be similar, as long as these children have not been taught probability theory of Bayes' rule.

What should matter, in contrast, is the ability to understand set inclusion, which is needed to recognize the proper set/subset relation of the sets $\{a\}$ and $\{a + c\}$. In this study, we used a test for understanding set/subset relations and test the additional hypothesis that its results should correlate with Bayesian intuitions.

Participants. Fifty-nine fifth-graders (mean age 11;4, range 9;10 to 12;6; 29 girls, 30 boys), 56 sixth-graders (mean age 12;4, range 10;10 to 13;5; 30 girls, 26 boys), and 53 seventh-graders (mean age 13;3, range 11;10 to 13;11; 30 girls, 23 boys) from a primary school in Osnabrück, a German city with about 160,000 inhabitants, participated in the study.

Materials and procedure. The children were tested in the second half of the school year during regular mathematics lessons. The teachers introduced the experimenter, who told the children that they would be asked unfamiliar problems, and encouraged them to solve as many as possible but to work carefully and write down every calculation, make drawings, or describe with words how they arrived at their answers. Children were rewarded with chocolate, and the teacher answered questions about the problems during the next mathematics lessons.

Each child was tested on eight Bayesian problems. Unlike in the previous studies, these problems had numerosities up to 100 (instead of up to 30). The problems were identical to the 10 problems used by Zhu and Gigerenzer (2006), apart from the omission of two to accommodate the time limit of 45 min, the duration of a standard school lesson. The appropriate time required was determined in a pilot study. Moreover, the two problems eliminated had an unfamiliar content for German children (e.g., reference to poor people without gloves for protection from cold weather). One other problem was adapted slightly to the German context, and a few numbers were modified so that the use of Bayes' rule produced a unique solution that differed from non-Bayesian rules. All eight problems are shown in the appendix. The last of these, the "cookies problem," contained a test of children's understanding of the relation between sets:

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 cookies, 24 are also round.

1. All round cookies are salty. Correct/Incorrect.
2. All salty cookies are round. Correct/Incorrect.
3. Most of the round cookies are salty. Correct/Incorrect.
4. Most of the salty cookies are round. Correct/Incorrect.
5. Most cookies are salty. Correct/Incorrect.
6. Most cookies are round. Correct/Incorrect.

Imagine you take out a pile of round cookies. How many of them are salty cookies? ___ of ___

The six questions test the ability to understand the relation between the sets H and D , here, salty and round cookies. Question 1 tests whether the child understands that round cookies are not a subset of salty cookies, that is, in general terms, that the set D is not a subset of the set H . Question 2 similarly tests the understanding that salty cookies are not a subset of round cookies, that is, that the set H is not a subset of the set D . Questions 3 and 4 resemble Questions 1 and 2, but test the understanding of relative quantity ("most"). Question 5 tests the understanding that salty cookies are the minority, that is, that the set H is smaller than the set $-H$. Finally, Question 6 tests the understanding that round cookies are also a minority, that is, that the set D is smaller than the set $-D$. The set inclusion question was asked only for the cookies problem.

The final question concerned quantitative Bayesian intuitions. For the cookies problem, Bayesian reasoning results in an expected proportion of 14 salty cookies out of every 38 round cookies in the pile. That is, the exact Bayesian response is 14 of 38. As in the previous studies, no other responses counted as Bayesian intuitions.

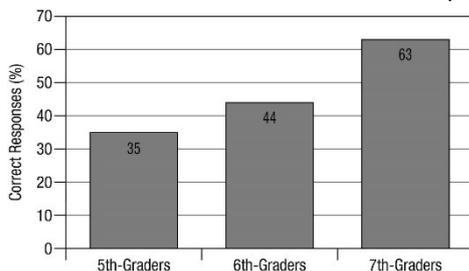
It has been suggested that dividing up the Bayesian question into two steps might improve Bayesian intuitions (Giroto & Gonzalez, 2008). To test this proposition, children were randomly assigned to two groups, one that received the quantitative question as stated above and a second that received the same question in two steps. In the two-step condition, the first question asked for the denominator, and the second for the numerator of the Bayesian posterior probability:

Imagine you take out a pile of round cookies.

How many of the cookies are round? ___

How many of them are salty cookies? ___

The order of the first seven problems was systematically varied to exclude order effects. The cookies problem with the additional questions was always presented last to avoid these set/subset questions having any influence on intuitive Bayesian reasoning. Children were also instructed not to return to the first seven questions after having answered the last problem.



Results

The average proportion of Bayesian intuitions did not differ between the two question formats, with an average of 3.7 ($SD = 3.0$) for single-step questions, and 3.8 ($SD = 3.1$) for the two-step question. Figure 10 thus shows the results aggregated over both

Figure 10. Bayesian intuitions among children in fifth, sixth, and seventh grade, averaged across eight problems. Correct responses are exact Bayesian responses.

groups. Bayesian intuitions show an increasing developmental trajectory, from 35% among fifth-graders to 44% among sixth-graders to 63% among seventh-graders, $F(2, 167) = 10.3, p < .001$. Note that these percentages refer to all eight problems; not every child could actually finish the set of problems within the time limit. If one uses as a base only the problems on which the children actually worked, the percentages are slightly higher for the younger children, 38% and 45%, and remain the same for the seventh-graders (63%).

In a next step, we investigated the individual differences within each age group. Out of all 168 children, 128 (76%) answered one or more problems correctly, whereas 40 (24%) could answer none (see Figure 11). One might assume that within each age group, the individual differences would be normally distributed around their average, but we had already seen that the distribution was highly left skewed among second-graders, and less so but nevertheless left skewed among fourth-graders. For fifth-graders, we still observed a left skewed distribution, with 18 children (31%) who could not solve a single problem and 19 children (32%) who could solve more than half (five to eight) of the problems. For sixth-graders, the number who could not solve a single problem remained the same, but those who solved more than half had increased, leaving a gap in the middle. For those in seventh grade, the distribution finally shifted its mode from zero to the maximum number of eight. As in Tables 1 and 2, individual differences are substantial, yet they appear to be u-shaped rather than symmetrically distributed around the mode (see Figure 11). These systematic age differences suggest a developmental shift from a mode of children without insight to a mode with insight within about two years rather than a gradual shift of the modal number of correct Bayesian intuitions.

Individual differences were associated with the understanding of set relations. Within each age group, the children who answered all six set-relation questions correctly (20, 27, and 29 children) had 20 to 30 percentage points more Bayesian intuitions than those who did not. This substantial effect is consistent with the numerical competency F2, which requires identifying the correct set and subset. Individual differences were also explained by the last report grade in mathematics: the better the grade, the more Bayesian responses, $\eta^2 = .22$ for fifth-graders, $\eta^2 = .27$ for sixth-graders, and $\eta^2 = .23$ for seventh-graders. In contrast, there were no gender differences in Bayesian intuitions: Boys achieved on average 5.3 ($SD = .87$) and girls 5.2 ($SD = .85$) correct responses. The ecological theory provides no reason to assume gender differences.

Discussion

The two main results of this study with natural frequencies are: (a) The average proportion of Bayesian intuitions is based not on a symmetric distribution but rather on a skewed one whose modes are either zero or the maximum number of correct answers. This change in the shape of distribution is consistent with Tables 1 and 2, where skewness increased the younger the age of the children. The distributions with the modes at the two extremes suggest an insight-like process where the performance resembles all or none rather than a value in the middle. (b) By seventh grade, children could solve the majority of problems, 63% of 424, with natural frequencies alone.

Consistent with the competency F2, the ability to understand set inclusion was found to explain part of the individual differences. When the present results are compared with those of the Chinese children in Zhu and Gigerenzer (2006), there is little evidence that Bayesian intuitions are limited to Chinese children with their superior math training. To make this comparison, we take the problems on which the German children actually worked as the basis because the Chinese children had no time limit. The Chinese fifth-graders (45 children in two experiments) solved 39% of the problems, compared with 38% in Study 4, which is as close a replication as can be expected. The Chinese sixth-graders (44 children) had 53.5% correct responses compared with 45% among the German children. The Chinese study had no seventh-graders, so the German group with 63% correct responses cannot be compared. All in all, any differences between the two studies can suggest only a small advantage of Chinese sixth-graders. Why similar results could not be found in the only other existing study, with Italian children, remains an open question: 75% of the Italian fourth- to sixth-graders could not solve a single problem, and even adults in this study could solve only 27% of the problems (Pighin et al., 2017), which is less than the fifth-graders in the present study.

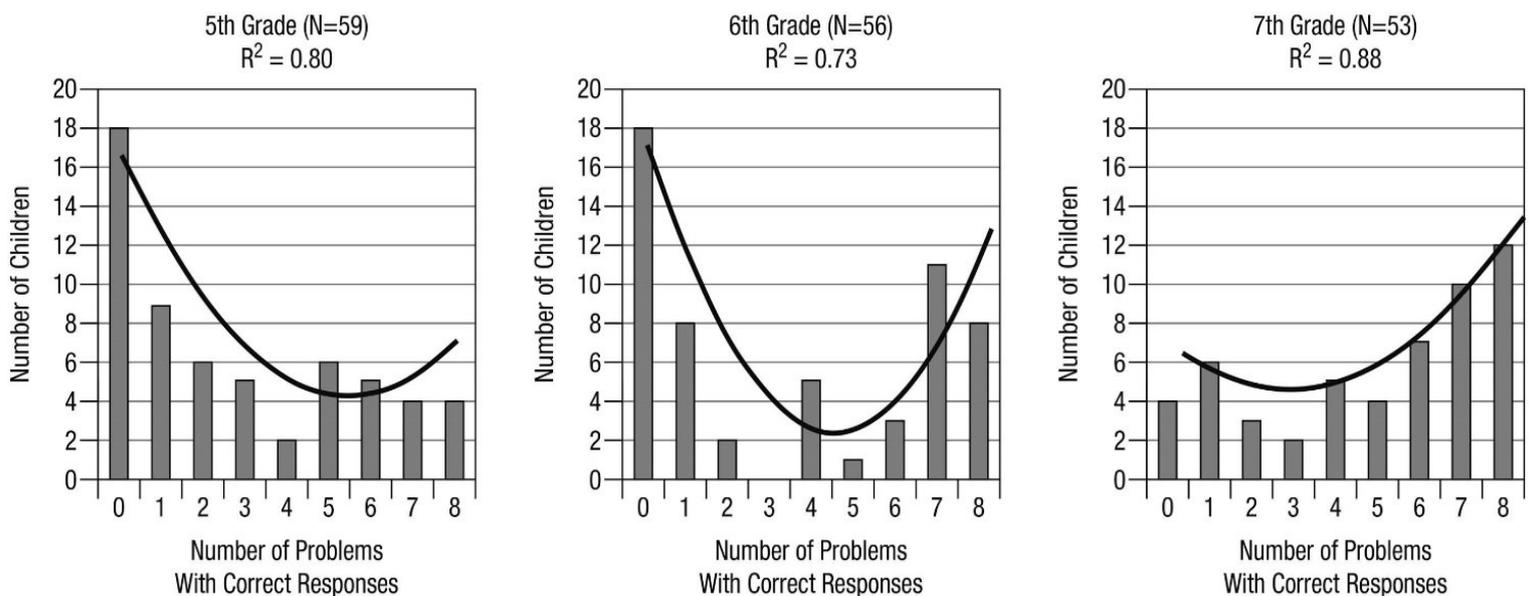


Figure 11. Individual differences within a grade appear to follow a U-shaped distribution, as opposed to a symmetric distribution around a mode. A two-degree polynomial whose minimum moves toward the right with increasing age can model this development (R^2 is the squared correlation between data and the polynomial). Note that the mode among the fifth- and sixth-graders is zero, whereas that of the seventh-graders is the maximum number of correct responses.

The comments and calculations the children wrote down and the questions asked revealed that a few children had to overcome additional challenges beyond the competencies F1 to F3. Some struggled with concepts such as *a group of people* or *a pile of cookies*, which we used as a proxy for *a random sample*, a term unfamiliar to most children of these age groups. For instance, several children asked how many cookies would fit in a hand.

In sum, Study 4 establishes that natural frequencies alone can reliably elicit Bayesian intuitions in fifth- to seventh-graders. This empirical result is consistent with the theoretical analysis that natural frequencies do not require understanding of probabilities, only of understanding and adding Arabic numerals, which most children at these ages have mastered.

Non-Bayesian Rules

The present studies also show that the younger children are, the less often they have Bayesian intuitions. In these cases, did children respond randomly or did they follow systematic non-Bayesian rules? We observed four systematic rules along with apparently random choices. To explain the systematic non-Bayesian rules we use the magic wand problem. Figure 12 shows the icon array of Figure 3 and the four non-Bayesian rules.

We analyzed all quantitative judgments in Study 2, that is, 546 judgments by second-graders and 520 by fourth-graders. We define the four non-Bayesian rules in terms of the types *a*, *b*, *c*, and *d*, as in Figure 2 (right-hand tree). The base rate is denoted as $e = a + b$. The total sample size is denoted as $t = a + b + c + d$. In the order of their frequency, the four non-Bayesian rules were:

1. *Hit-rate rule*: a/e , that is, $\text{count}\{a\text{-tokens}\}/\text{count}\{e\text{-tokens}\}$
2. *Denominator-error rule*: a/c , that is, $\text{count}\{a\text{-tokens}\}/\text{count}\{c\text{-tokens}\}$
3. *Base-rate rule*: e/t , that is, $\text{count}\{e\text{-tokens}\}/\text{count}\{\text{all tokens}\}$
4. *Evidence-rate rule*: $(a + c)/t$, that is, $\text{count}\{a\text{-tokens}\}/\text{count}\{\text{all tokens}\}$

The first rule, a/e , is called the *hit-rate rule* because it corresponds to the probability $p(D|H)$, also known as the hit rate. It generates the estimate “four out of five” instead of “four out of 16.” Among second-graders, 16% and 22% of all problems were answered in this way for icons and natural frequencies, respectively. Among fourth-graders, the corresponding numbers declined to 12% and 17%. The hit-rate rule has also been documented in adults (Gigerenzer & Hoffrage, 1995). It also has been called “representative thinking” (Dawes, 1986).

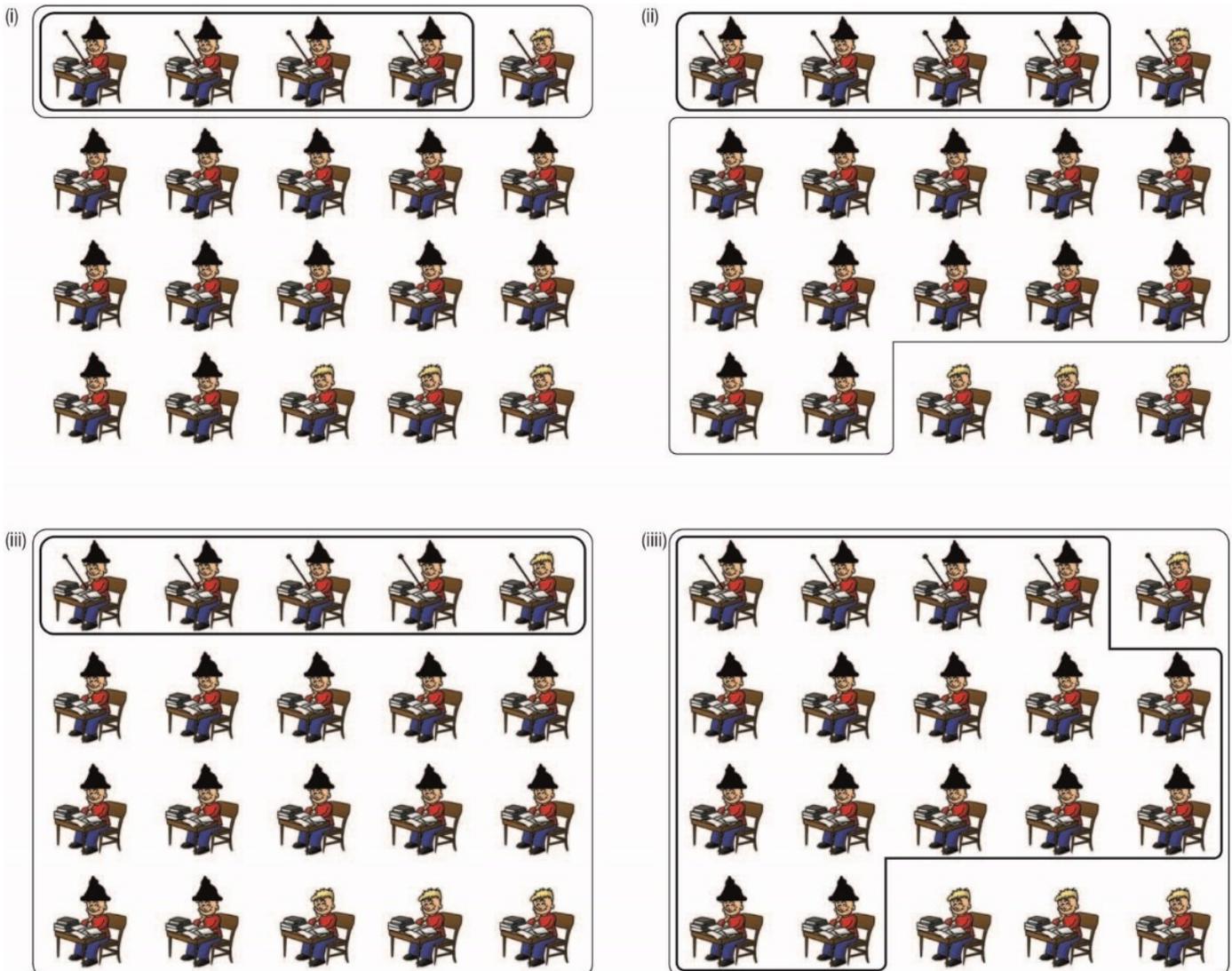


Figure 12. Illustration of the four non-Bayesian rules identified, using the icon array in Figure 3. These are (i) a/e (hit-rate rule), (ii) a/c (denominator-error rule), (iii) e/t (base-rate rule), and (iiii) $(a + c)/t$ (evidence-rate rule). For each of children’s rules, the thick line in the icon array indicates the numerator and the thinner line indicates the denominator.

The second rule, a/c , the *denominator error*, generates the estimate “four out of 12” in the magic wand problem. Among second-graders, 11% and 7% of all problems were answered in this way for icon arrays and natural frequencies, respectively. Among fourth-graders, these numbers were 5% and 8%, respectively. This rule has not been documented in adults. Like the hit-rate rule, it gets the numerator right but not the denominator.

The *base-rate rule* confuses the base rate e/t with the posterior probability. It leads to the answer “five out of 20,” which is the base rate of magic wands. Among second-graders, 5% and 12% of all problems were answered in this way for icon arrays and natural frequencies, respectively. Among fourth-graders, these numbers declined to 4% and 5%, respectively. The base-rate rule has been previously reported in adults (Gigerenzer & Hoffrage, 1995) and has also been termed *conservatism*, reflecting that one relies only on the base rate and ignores the diagnostic information.

The fourth non-Bayesian rule was $(a + c)/t$, which leads to the answer “16 out of 20.” Among second-graders, 4% and 4% of all problems were answered in this way for icon arrays and natural frequencies, respectively. Among fourth-graders, these numbers were 2% and 7%. The rule corresponds to the probability $p(D)$, that is, the unconditional probability of data, such as the probability of wearing a magic hat. Its logic contrasts with the base-rate rule because it ignores the base rate and considers the evidence rate only. This *evidence-rate rule* has not been reported among adults (Gigerenzer & Hoffrage, 1995).

Together with the Bayesian intuitions, the four rules account for 57% and 81% of all responses of the second- and fourth-graders, respectively, averaged across icon array and natural frequency conditions. The remaining responses could not be identified, or when they could, their frequency was so small that they might easily have arisen from random choice.

In sum, children appear to rely on four non-Bayesian rules. The rules put two of the numbers into relation, except for the least frequent one, evidence-rate, where three numbers are put into relation. Three of the rules correspond to basic statistical concepts: $p(D|H)$, $p(H)$, and $p(D)$. The frequency of these non-Bayesian rules is generally lower with icon arrays and also decreases with age.

We can now understand the kind of errors children make, of which there appear to be three types: errors of identification, combining, and counting. First, at the most fundamental level, children may fail to identify the correct types (subsets). An example is the hit-rate rule, where children rely on types a and e instead of a and c . Second, some children are able to identify the correct types but fail to combine them according to Equation 5. An example is the denominator error, where a and c are correctly identified, but combined as a/c instead of $a/(a + c)$. Third, other children are able to identify the correct types and combine them correctly, but make counting errors. Tables 1 and 2 provide estimates for the frequency of these counting errors.

General Discussion

In this article, we have provided a general framework for understanding Bayesian intuitions in children untutored in probability theory or Bayes' rule. The key theoretical idea is that the numerical competencies required by Bayesian intuitions depend on the external representation of numerosity. In a series of experiments, we showed that icon arrays can elicit Bayesian intuitions in children as young as second-graders. These children untutored in probability theory could find the exact Bayesian response for 22% to 32% of all problems, and fourth-graders for 54% to 61% of problems. We also report for the first time that icon arrays can elicit Bayesian intuitions in children diagnosed with developmental dyscalculia, and at a level approximating other children of the same age. The theoretical analysis and the supporting empirical results indicate that children's numerical reasoning has more potential than assumed so far, but that this potential needs to be tapped with the help of appropriate external representations.

One might object that the children were not completely unfamiliar with Bayesian problems because we used an example problem to explain the task, as described above. Without any instructions, however, children as young as second-graders would not have understood what the task was about. Note that the instructions were limited to the example problem (which was not included in the test set), and no general principle was explained. If we had systematically taught children how to solve Bayesian problems, something we suggest in the section Teaching Bayesian Reasoning (see below), performance would likely have improved.

In conclusion, we extend these findings to propose a solution to three apparent contradictions in the literature on Bayesian reasoning in humans and animals. We then derive the implications of the ecological approach for teaching Bayes' rule and probability in general.

Three Puzzles in the Literature Revisited

The role of Bayes' rule in contemporary theories of cognition can be characterized as one of both unification and fragmentation. On the one hand, it has been proposed as a unifying framework for theories of cognition in humans and other animals, bringing perception, memory, reasoning, and learning together into a single framework: the mind as a Bayesian (e.g., Chater, Tenenbaum, & Yuille, 2006; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Schooler & Anderson, 1997). The same hope is expressed in the neuroscience vision of the “Bayesian brain” (Friston, 2010). On the other hand, Bayes' rule has fragmented psychology into those who view the mind as a Bayesian and those who conclude that the mind systematically fails to reason according to Bayes' rule (Kahneman, 2011; Kahneman & Tversky, 1972). Yet the divide goes deeper. Whereas Anderson, Chater, Friston, and others propose that the fast and automatic processes of perception, attention, and memory are Bayesian inferences, Kahneman and others argue that the same automatic processes (called “System 1”) are incompatible with the rules of probability. This fundamental divide within psychology is often not noticed, in part because the mind-as-a-Bayesian view is virtually never cited in the vast literature on “System 1.”

The fragmentation has also led to three puzzling contradictions (Hertwig, 2015; Mandel, 2014; Schulze & Hertwig, 2018). We briefly review the three and suggest how the proposed ecological theory can resolve their apparently contradictory results.

The 1970s puzzle in Bayesian reasoning. The 1970s puzzle is the question why people (adults) were reported before 1970 to be approximate, albeit conservative Bayesians, and after 1970 as lacking Bayesian reasoning and neglecting base rates, which is the opposite of conservatism.

In the 1960s, Ward Edwards and colleagues (inspired by Rouanet, 1961) set up the first experimental program on Bayesian

reasoning in adults and concluded that people are approximate Bayesians, albeit conservative: “opinion change is very orderly, and usually proportional to numbers calculated from Bayes’ theorem” (Edwards, 1968, p. 17). *Conservative* here refers to overly weighting base rates. Starting in the early 1970s, Kahneman and Tversky (1972) set up their experimental program on systematic biases and concluded: “In his evaluation of evidence, man is apparently not a conservative Bayesian: he is not Bayesian at all.” They reported that people give too little weight to base rates, later known as the *base-rate fallacy*. This result was widely hailed. “The genuineness, the robustness, and the generality of the base-rate fallacy are matters of established fact” (Bar-Hillel, 1980, p. 215). Von Winterfeldt and Edwards (1986) struggled for an explanation of the puzzling reversal of facts around 1970. They tentatively attributed the problem to the fact that people learn arithmetic but not probability theory (pp. xiii–xiv). Yet that does not explain why people were approximate Bayesians during the 1960s and not so in the 1970s. The tasks differ in several ways, such as that those from Edwards were relative likelihood tasks while many of Kahneman & Tversky’s were judgments of proportions (Gallistel et al., 2014). Here we focus on the representation of information.

A closer look at the experimental task provides a clue. Hertwig (2015; Schulze & Hertwig, 2018) noted that the difference corresponds to that between decisions from experience and decisions from description, which often lead to diverging results. The present ecological theory can explain why. Edwards and colleagues used a probability learning paradigm similar to natural sampling. In contrast, the studies on adults after 1970 typically do not involve sequential learning but instead rely on tasks corresponding to the conditional probability representation of the magic wand problem above. When one learns through natural sampling, the numerical competencies required for Bayesian reasoning correspond to F1 to F3, whereas conditional probability tasks, such as the cab problem (Tversky & Kahneman, 1980), require competencies P1 to P3. This explanation can be tested by giving the same people the same problems but in the two different paradigms. For instance, Armstrong, Spaniol, and Persaud (2018) tested 50 physicians at St. Michael’s Hospital in Toronto in two formats: direct experience of 100 patient cases and a conditional probability format (relative frequencies) in an intraindividual design. The mean absolute error of the positive predictive value was 16 percentage points in the natural sampling condition, but 33 in the relative frequency condition. In general, several studies have showed that with natural sampling or natural frequency, adults succeed in reasoning the Bayesian way more often than with conditional probabilities or relative frequencies (e.g., Betsch, Biel, Eddelbuttel, & Mock, 1998; Hoffrage, Hafenbrädl, & Bouquet, 2015; McDowell & Jacobs, 2017). Thus, the possible answer to the 1970s puzzle is not a decline in people’s competencies but a change in the experimental set-up: from testing inferences based on natural sampling to testing inferences based on conditional probability representations, which implies different cognitive competencies.

The phylogenetic puzzle. The phylogenetic puzzle is the question why honey bees, bumblebees, birds, and other animal species exhibit approximate Bayesian reasoning, but human adults fail to do so.

Animal researchers began around the 1980s to study the question whether animal behavior is consistent with Bayes’ rule. For instance, squirrels have been reported to make decisions about whether to stay or to leave a patch according to Bayes’ rule (McNamara & Houston, 1985). Similarly, Bayesian reasoning was reported for the domestic pigeon (Real, 1991), for three-spined sticklebacks’ group decision making, and for female animals searching for suitable males (Pérez-Escudero & de Polavieja, 2011). In addition, many animal species have been reported to be highly sensitive to changes in frequency distributions in their environments (Gallistel, 1990; Hanus & Call, 2014; Real & Caraco, 1986). In their review, McNamara, Green, and Olsson (2006, p. 243) conclude that “we should often expect to see ‘Bayesian-like’ decision-making in nature.”

These results also contrast with the findings in the 1970s that human adults systematically deviate from Bayesian reasoning. Is there a phylogenetic reversal in Bayesian reasoning from humans to other animals? We do not think so. Rather, the solution to this apparent puzzle likely lies in the representation of numerosity.

Animal studies typically use a probability learning paradigm, where joint events are learned, such as cue (yes/no) and food (yes/no). This way of learning is also known as natural sampling, whose result corresponds to the natural frequencies in Figure 2 and the icon array in Figure 3. That is, the tasks animals are confronted with require different numerical competencies than the ones to which humans in the heuristics-and-biases program are subjected. Humans need to understand and reason using conditional probabilities (competencies P1 to P3), whereas animals only need to count and keep a current tally, that is, approximate competencies I1 to I3. This ability to count and keep an approximate tally has been documented for various animal species (Boysen & Capaldi, 1993).

Thus, we propose that the phylogenetic puzzle is not really about differences between species but rather about differences in external representations of numerosity. As a test of this explanation, one can put humans in the same experimental situation in which animals are studied. When humans are tested in a natural sampling paradigm, base-rate neglect largely disappears and they can easily match the performance of animals in Bayesian reasoning (Betsch et al., 1998). This result is consistent with the finding that human frequency processing is fairly accurate under naturalistic conditions (Zacks & Hasher, 2002). That is, the problem is not in the human mind, but in the representation of numerosity. In this way, the phylogenetic puzzle can be resolved.

The ontogenetic puzzle. The ontogenetic puzzle has been expressed by Alison Gopnik (2014): “Why are grown-ups often so stupid about probabilities when even babies and chimps can be so smart?” This puzzle is more general than about Bayesian reasoning alone.

Studies on infants’ intuitive statistics have concluded a surprising capability for a number sense. For instance, 6-month-old infants have been reported to differentiate between sets of disks, provided that the sets differ by a large ratio (specifically, between eight and 16 but not between eight and 12 disks; Xu & Spelke, 2000). Moreover, infants are able to use proportions to predict outcomes of random draws (Denison & Xu, 2010; Xu & Garcia, 2009). Once again, a closer look at the experimental tasks reveals the same kind of difference as between animals and humans, and between adults before and after the 1970s. Unlike adults in studies since the 1970s, infants typically learn about probabilities by experience, that is, by sequential updating of information, such as seeing red and white balls being drawn one by one from an urn. In

fact, the experiments with infants often use countable objects that resemble icon arrays (Denison & Xu, 2019). This corresponds to the book-bag and poker-chips tasks used in the 1960s by Edwards and colleagues, where one ball after another is drawn from an urn.

All three puzzles involve an apparent reversal of a developmental progression—smart animals, smart babies, and smart adults before the 1970s, and not-so-smart adults thereafter. The present ecological analysis provides a common explanation for all three puzzles. If animals, infants, or adults learn about probabilities from direct experience by natural sampling—or are presented with its final tally in form of icon arrays or natural frequencies—then performance tends to be high. If humans are given the same information in terms of conditional probabilities, performance is comparably low.

Teaching Bayesian Reasoning

In this article, we have not dealt with systematically teaching Bayesian reasoning, only with the untutored mind. Learning statistical thinking, however, equips children with an essential decision-making skill for their personal and professional lives. The failure to reason the Bayesian way has caused innumerable serious, even deadly errors in legal and medical contexts (see, e.g., Dawid, 2002; Gigerenzer, 2014; Good, 1996) and continues to be a problem; many legal and medical professionals to the present day have never learned how to reason with probabilities (Gigerenzer, 2014; Lindsey, Hertwig, & Gigerenzer, 2003).

Lack of Bayesian reasoning, or of statistical reasoning in general, has sometimes been attributed to a fundamental design flaw in the human mind, which is allegedly not built “to work by the rules of probability” (Gould, 1992, p. 469). Similar to genetic theories of dyscalculia, this view implies that little can be done to improve the situation. The present ecological theory, in contrast, implies how to make the apparent flaw largely disappear. This can be done by simply using external representations of numerosity matched to the competencies of individuals, as argued in this article, and also by using these insights to systematically teach Bayesian reasoning.

The key idea of an ecological approach is to teach individuals how to translate probabilities into a representation they can master, such as icon arrays or natural frequencies. In medicine, where the lack of Bayesian reasoning can lead to harmful outcomes, such programs have been tested. In a continuous medical education (CME) study with 160 gynecologists, only 21% could identify the correct posterior probability that a woman with a positive screening mammogram actually has breast cancer. This was slightly worse than chance (25%), given four alternatives to choose from, ranging from 1% to 90%. After a short training in how to translate conditional probabilities into natural frequencies, this number increased to 87% (Gigerenzer et al., 2007). Among 104 final-year medical students who were given the same task, only 23% (again, slightly below chance) could determine the correct posterior probability, whereas after a 90-min training that included natural frequencies, 88% were able to translate conditional probabilities into natural frequencies and correctly estimate the resulting posterior probability (Jenny, Keller, & Gigerenzer, 2018). Meanwhile, *natural frequencies* have become a technical term in evidence-based medicine, and promoted by several medical organizations for patient– doctor communication (e.g., Akl et al., 2011; McDowell & Jacobs, 2017; National Health Service, 2017).

The most efficient way to teach Bayesian reasoning, however, would be to start much earlier, in school.

Improving Teaching of Statistics

Statistics education has become part of the high school and university curricula in many countries. However, teaching efforts are targeted at calculating with probabilities. Even when the information is represented in frequencies, students are instructed to solve the task with probabilities (Weber et al., 2018). This practice has been shown to create a mental set (“Einstellung”) effect (Luchins, 1942), in which students try to rigidly apply a previously learned solution and are “blind” to an easier one. An example from the present study is training people to apply Equation 2 when Equation 4 would lead to a quicker answer and better understanding. In a study with university students enrolled in a teaching math program, about half of the participants translated natural frequencies into probabilities (rather than calculating with natural frequencies), which led to substantial decreases in Bayesian solutions compared with those who did not (Weber et al., 2018). Only 9% and 12% of students who translated natural frequencies into probabilities solved the two problems, respectively, whereas 75% and 41% of those who calculated with natural frequencies did so. Similarly, when the information was given in probabilities, those who calculated with probabilities provided only 12% to 15% correct solutions, whereas those who translated probabilities into natural frequencies provided 45% to 67% correct ones. The authors speak of a “frequency phobia” among students who are trained in the probability calculus, which maintains their “probability blindness” rather than helping them to understand. In this way, the positive effect of natural frequencies is counteracted and underestimated in studies with adults trained in probability (as opposed to untutored children).

Butterworth (2001) suggested that current educational practices are partly responsible for the observed inability of adults to reason the Bayesian way. The present study suggests a revision of practice: to teach Bayesian thinking first with icon arrays that require counting alone to children in second to fourth grade, then introduce natural frequencies that require adding Arabic numerals, and only thereafter introduce probabilities that require competencies in adding and multiplying numerical probabilities. In this way, students can early acquire confidence that they are able to solve Bayesian problems and, if they fail with a more demanding representation, can always go back to one that they have already mastered.

Long-Term Effects of Teaching

Teaching Bayesian reasoning may have two goals: teaching for the test and teaching for life. The first is measured by the short-term performance after teaching, the second by the long-term retention of what has been learned. German psychology students were trained in Bayes’ rule using either conditional probabilities, natural frequency trees, or icon arrays (Sedlmeier & Gigerenzer, 2001). The immediate learning effect was higher for natural frequency trees and icon arrays (from 10% Bayesian responses before training to about 90% after) than for probabilities (from 0% to about 65%). More important, 5 weeks after training, the perfor-

mance of the groups who had learned to use natural frequencies and icon arrays remained at a high 90% when tested with new problems, whereas that of the group who were trained using conditional probabilities dropped to 20%. A replication study with psychology students in the US found similar results: Four weeks after training, the performance of the group trained with natural frequency trees or icon arrays dropped only slightly or remained the same, whereas that of the group trained with conditional probabilities had deteriorated from 49% immediately after learning to 6% (Ruscio, 2003). Studies with 10th- and 13th-graders, who in part had received training in probability theory, showed that a computerized probability training led to an average of 35% and 50% Bayesian solutions, respectively, whereas a training in natural frequency trees resulted in 60% and 80%, respectively (Wassner, 2004). The loss of performance following training was again steeper for probability than for natural frequency formats (Wassner, Martignon, & Biehler, 2004).

Textbooks typically teach Bayes' rule in the form of Equation 2 or a probability tree as the left-hand one in Figure 2. With this conventional representation, only 10% of 11th-graders (aged 16–18) could find the Bayesian solution. Yet when the pupils encountered natural frequency trees, despite unfamiliarity with them, about 51% solved the problem (Binder, Krauss, Bruckmaier, & Marienhagen, 2018). In reaction to this study, the Ministry of Education in Bavaria, Germany, has introduced as of 2018 natural frequencies into the math curriculum to improve teaching of Bayes' rule (Staatsinstitut für Schulqualität und Bildungsforschung München, 2018).

The general lesson is that teaching statistics should match the representation of numerosity to the numerical competencies of the learner. In this way, teaching can exploit the intuitive abilities of young children in counting and handling Arabic numbers (as defined in I1 to I3 and F1 to F3). This ecological perspective can boost performance and spare both children and adults the frustrating experience of failing to learn Bayes' rule with the "help" of probability trees.

An Ecological Perspective for the Development of Probabilistic Thinking

This article has dealt with an elementary form of Bayesian reasoning that involves a binary criterion and a binary cue. Whether icon arrays can elicit Bayesian intuitions in children faced with more complex tasks, such as with more than one cue or more than two cue values, is unknown. In adults, two studies showed that natural frequencies, with or without visualization, can elicit Bayesian reasoning for problems with two cues (Binder et al., 2018) and up to three cues, three cue values, and three hypotheses (Hoffrage, Krauss, Martignon, & Gigerenzer, 2015). These complex tasks did not decrease the facilitating effect of natural frequencies, and a short instruction boosted performance to 73% (three cue values) and 81% (two cues) Bayesian responses (Hoffrage et al., 2015). Thus, icon arrays and natural frequencies may be able to facilitate Bayesian intuitions for a much broader class of problems than studied in this article.

The ecological perspective emphasizes that mathematically equivalent representations are not generally psychologically equivalent. This view may also be of relevance for the general issue of framing numerical information. Logically equivalent frames are not necessarily psychologically equivalent (McKenzie & Nelson, 2003; Sher & McKenzie, 2006), and finding a proper frame can be a key to discovery and success. Mathematics and physics are prime examples of disciplines that value the importance of notation and frame. Feynman (1965) conjectured that two logically equivalent representations of the same formula or the same physical law can evoke different mental pictures and assist in making discoveries: "psychologically they are different because they are completely unequivalent when you are trying to guess new laws" (p. 53). Feynman put the power of representations into work by means of his diagrams.

The ecological theory proposed here is based on the view that thinking is shaped by the match between cognitive processes and the external representation of information. This perspective applies to statistical reasoning in general. Using this theory, we could show that children untutored in probability theory and even with dyscalculia have Bayesian intuitions. Thus, it seems that Laplace (1814/1951, p. 196) was right after all when he wrote that probability is "only common sense reduced to calculus."

Context of the Research

This research has an unusual history. It was motivated by the first and last authors' experience in teaching hundreds of physicians risk literacy in CME. For instance, when we asked gynecologists to estimate the probability that a woman has breast cancer if she had a positive screening mammogram, their estimates varied widely between 1% and 90%. In general, the problem was that most doctors were unable to infer the Bayesian posterior probability from information about the base rate, sensitivity, and false positive rate. When we taught physicians to translate these conditional probabilities into natural frequencies, however, most immediately understood the correct answer. That result demonstrated how important the representation of numerosity is for understanding. Leading from there, we hypothesized that even young children might be able to solve Bayesian problems if a representation could be found that only requires the competences that children already have. Icon arrays are one such representation, where basically all one needs to know is how to count and what to count. At the schools where we conducted the experiments, teachers reacted with disbelief, expecting their fourth-graders, not to speak of second-graders, to be entirely baffled by these tasks. This disbelief was even stronger when we approached teachers and clinical psychologists to test children diagnosed with dyscalculia. To the teachers' surprise, however, a substantial proportion of children could determine the exact Bayesian answer, suggesting new ideas about the reasons for dyscalculia and about improved teaching methods. This article is the story of a simple idea, that statistical thinking depends on the external representation of numerosity, which can improve teaching of statistical thinking at the primary and secondary level and in schools of medicine and law. The idea is to teach statistical thinking with icon arrays first, before moving on to frequencies and probabilities.

References

- Akl, E. A., Oxman, A. D., Herrin, J., Vist, G. E., Terrenato, I., Sperati, F., . . . Schünemann, H. (2011). Using alternative statistical formats for presenting risks and risk reductions. *Cochrane Database of Systematic Reviews*, 3, CD006776. <http://dx.doi.org/10.1002/14651858.CD006776.pub2>
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

- Antell, S. E., & Keating, D. P. (1983). Perception of numerical invariance in neonates. *Child Development*, 54, 695–701. <http://dx.doi.org/10.2307/1130057>
- Armstrong, B., Spaniol, J., & Persaud, N. (2018). Does exposure to simulated patient cases improve accuracy of clinicians' predictive value estimates of diagnostic test results? A within-subjects experiment at St Michael's Hospital, Toronto, Canada. *British Medical Journal Open*, 8, e019241. <http://dx.doi.org/10.1136/bmjopen-2017-019241>
- Artelt, C., Demmrich, A., & Baumert, J. (2001). Selbstreguliertes lernen. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann, & M. Weiß (Eds.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 71–298). Opladen, Germany: Leske + Budrich. http://dx.doi.org/10.1007/978-3-322-83412-6_8
- Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, 44, 211–233. [http://dx.doi.org/10.1016/0001-6918\(80\)90046-3](http://dx.doi.org/10.1016/0001-6918(80)90046-3)
- Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53, 370–418.
- Beddington, J., Cooper, C. L., Field, J., Goswami, U., Huppert, F. A., Jenkins, R., . . . Thomas, S. M. (2008). The mental wealth of nations. *Nature*, 455, 1057–1060. <http://dx.doi.org/10.1038/4551057a>
- Betsch, T., Biel, G. M., Eddelbittel, C., & Mock, A. (1998). Natural sampling and base-rate neglect. *European Journal of Social Psychology*, 28, 269–273. [http://dx.doi.org/10.1002/\(SICI\)1099-0992\(199803/04\)28:2269:AID-EJSP872>3.0.CO;2-U](http://dx.doi.org/10.1002/(SICI)1099-0992(199803/04)28:2269:AID-EJSP872>3.0.CO;2-U)
- Binder, K., Krauss, S., Bruckmaier, G., & Marienhagen, J. (2018). Visualizing the Bayesian 2-test case: The effect of tree diagrams on medical decision making. *PLoS ONE*, 13, e0195029. <http://dx.doi.org/10.1371/journal.pone.0195029>
- Boysen, S. T., & Capaldi, E. J. (Eds.). (1993). *The development of numerical competence: Animal and human models*. Hillsdale, NJ: Erlbaum.
- Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychonomic Bulletin & Review*, 15, 284–289. <http://dx.doi.org/10.3758/PBR.15.2.284>
- Brase, G. (2009). Pictorial representations and numerical representations in Bayesian reasoning. *Applied Cognitive Psychology*, 23, 369–381. <http://dx.doi.org/10.1002/acp.1460>
- Brase, G. L. (2014). The power of representation and interpretation: Doubling statistical reasoning performance with icons and frequentist interpretations of ambiguous numbers. *Journal of Cognitive Psychology*, 26, 81–97. <http://dx.doi.org/10.1080/20445911.2013.861840>
- Brunswik, E. (1939). Probability as a determiner of rat behavior. *Journal of Experimental Psychology*, 25, 175–197. <http://dx.doi.org/10.1037/h0061204>
- Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62, 193–217. <http://dx.doi.org/10.1037/h0047470>
- Butterworth, B. (2001). Statistics: What seems natural? *Science*, 292, 853–855. <http://dx.doi.org/10.1126/science.292.5518.853c>
- Butterworth, B. (2005). The development of arithmetical abilities. *Journal of Child Psychology and Psychiatry*, 46, 3–18. <http://dx.doi.org/10.1111/j.1469-7610.2004.00374.x>
- Carey, S. (2009). *The origin of concepts*. New York, NY: Oxford University Press. <http://dx.doi.org/10.1093/acprof:oso/9780195367638.001.0001>
- Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10, 287–291. <http://dx.doi.org/10.1016/j.tics.2006.05.007>
- Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision Making*, 7, 25–47.
- 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. [https://doi.org/10.1016/0010-0277\(95\)00664-8](https://doi.org/10.1016/0010-0277(95)00664-8)
- Dantzig, T. (1954). *Number: The language of science*. New York, NY: MacMillan.
- Daston, L. J. (1988). *Classical probability in the Enlightenment*. Princeton, NJ: Princeton University Press.
- Dawes, R. M. (1986). Representative thinking in clinical judgment. *Clinical Psychology Review*, 6, 425–441. [https://doi.org/10.1016/0272-7358\(86\)90030-9](https://doi.org/10.1016/0272-7358(86)90030-9)
- Dawid, A. P. (2002). Bayes's theorem and weighting evidence by juries. *Proceedings of the British Academy*, 113, 71–90.
- Dehaene, S. (2011). *The number sense: How the mind creates mathematics* (2nd revised ed.). New York, NY: Oxford University Press. (Original work published 1997)
- Denison, S., & Xu, F. (2010). Twelve- to 14-month-old infants can predict single-event probability with large set sizes. *Developmental Science*, 13, 798–803. <http://dx.doi.org/10.1111/j.1467-7687.2009.00943.x>
- Denison, S., & Xu, F. (2019). Infant statisticians: The origins of reasoning under uncertainty. *Perspectives on Psychological Science*, 14, 499–509. <http://dx.doi.org/10.1177/1745691619847201>
- Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17– 52). New York, NY: Wiley.
- Feynman, R. P. (1965). *The character of physical law*. Cambridge, MA: MIT Press.
- Fischbein, E. (1975). *The intuitive sources of probabilistic thinking in children*. Dordrecht, the Netherlands: D. Reidel. <https://doi.org/10.1007/978-94-010-1858-6>
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11, 127–138. <http://dx.doi.org/10.1038/nrn2787>
- Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology*, 28, 210–216. <http://dx.doi.org/10.1037/a0014474>
- Galesic, M., Gigerenzer, G., & Straubinger, N. (2009). Natural frequencies facilitate accurate judgments about medical screenings for elderly and people with lower numeracy skills. *Medical Decision Making*, 29, 368– 371. <http://dx.doi.org/10.1177/0272989X08329463>
- Gallistel, C. R. (1990). *The organization of learning*. Cambridge, MA: MIT Press.
- Gallistel, C. R., Krishan, M., Liu, Y., Miller, R., & Latham, P. E. (2014). The perception of probability. *Psychological Review*, 121, 96–123. <http://dx.doi.org/10.1037/a0035232>
- Gelman, R., & Gallistel, C. R. (1978). *The child's understanding of number*. Cambridge, MA: Harvard University Press.
- Gelman, R., & Tucker, M. F. (1975). Further investigations of the young child's conception of number. *Child Development*, 46, 167–175. <http://dx.doi.org/10.2307/1128845>
- Gigerenzer, G. (1992). Discovery in cognitive psychology: New tools inspire new theories. *Science in Context*, 5, 329–350. <http://dx.doi.org/10.1017/S0269889700001216>
- Gigerenzer, G. (2014). *Risk savvy: How to make good decisions*. New York, NY: Viking.
- Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients to make sense of health statistics. *Psychological Science in the Public Interest*, 8, 53–96. <http://dx.doi.org/10.1111/j.1539-6053.2008.00033.x>
- Gigerenzer, G., & Goldstein, D. G. (1996). Mind as computer: Birth of a metaphor. *Creativity Research Journal*, 9, 131–144. http://dx.doi.org/10.1207/s15326934crj0902&3_3
- Gigerenzer, G., Hertwig, R., & Pachur, T. (Eds.). (2011). *Heuristics: The foundations of adaptive behavior*. New York, NY: Oxford University Press. <http://dx.doi.org/10.1093/acprof:oso/9780199744282.001.0001>

- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102, 684–704. <http://dx.doi.org/10.1037/0033-295X.102.4.684>
- Gigerenzer, G., & Hoffrage, U. (1999). Overcoming difficulties in Bayesian reasoning: A reply to Lewis & Keren and Mellers & McGraw. *Psychological Review*, 106, 425–430. <http://dx.doi.org/10.1037/0033-295X.106.2.425>
- Gigerenzer, G., & Hoffrage, U. (2007). The role of representation in Bayesian reasoning: Correcting common misconceptions. *Behavioral and Brain Sciences*, 30, 264–267. <http://dx.doi.org/10.1017/S0140525X07001756>
- Gigerenzer, G., & Murray, D. J. (2015). *Cognition as intuitive statistics*. New York, NY: Psychology Press. <https://doi.org/10.4324/9781315668796>
- Gilmore, C. K., McCarthy, S. E., & Spelke, E. S. (2007). Symbolic arithmetic knowledge without instruction. *Nature*, 447, 589–591. <http://dx.doi.org/10.1038/nature05850>
- Giroto, V., & Gonzalez, M. (2008). Children's understanding of posterior probability. *Cognition*, 106, 325–344. <http://dx.doi.org/10.1016/j.cognition.2007.02.005>
- Good, I. J. (1996). When batterer becomes murderer. *Nature*, 381, 481. <http://dx.doi.org/10.1038/381481a0>
- Gopnik, A. (2014, January 10). The surprising probability gurus wearing diapers. *The Wall Street Journal*. Retrieved from <https://www.wsj.com/articles/the-surprising-probability-gurus-wearing-diapers-1389401010>
- Gould, S. J. (1992). *Bully for brontosaurus: Further reflections in natural history*. New York, NY: Penguin Books.
- Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14, 357–364. <http://dx.doi.org/10.1016/j.tics.2010.05.004>
- Hacking, I. (1975). *The emergence of probability*. Cambridge, UK: Cambridge University Press.
- Hafenbrädl, S., & Hoffrage, U. (2015). Toward an ecological analysis of Bayesian inferences: How task characteristics influence responses. *Frontiers in Psychology*, 6, 939. <http://dx.doi.org/10.3389/fpsyg.2015.00939>
- Hanus, D., & Call, J. (2014). When maths trumps logic: Probabilistic judgements in chimpanzees. *Biology Letters*, 10, 20140892. <http://dx.doi.org/10.1098/rsbl.2014.0892>
- Hertwig, R. (2015). Decisions from experience. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (Vol. 1, pp. 239–267). Chichester, UK: Wiley Blackwell. <http://dx.doi.org/10.1002/9781118468333.ch8>
- Hoemann, H. W., & Ross, B. M. (1982). Children's concepts of chance and probability. In C. J. Brainerd (Ed.), *Children's logical and mathematical cognition: Progress in cognitive development research* (pp. 93–121). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4613-9466-2_3
- Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, 73, 538–540. <http://dx.doi.org/10.1097/00001888-199805000-00024>
- Hoffrage, U., Hafenbrädl, S., & Bouquet, C. (2015). Natural frequencies facilitate diagnostic inferences of managers. *Frontiers in Psychology*, 6, 642. <http://dx.doi.org/10.3389/fpsyg.2015.00642>
- Hoffrage, U., Krauss, S., Martignon, L., & Gigerenzer, G. (2015). Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. *Frontiers in Psychology*, 6, 1473. <http://dx.doi.org/10.3389/fpsyg.2015.01473>
- Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Medicine. Communicating statistical information. *Science*, 290, 2261–2262. <http://dx.doi.org/10.1126/science.290.5500.2261>
- Ibrah, G. (2000). *A universal history of numbers*. New York, NY: Wiley.
- Jacobs, C., & Petermann, F. (2005). Aufmerksamkeitsstörungen im Kindesalter: Konzept und Wirksamkeit des ATTENTIONER-Programms [Attention deficiency disorders in children: The concept and efficacy of the ATTENTIONER program]. *Verhaltenstherapie & Verhaltensmedizin*, 26, 317–341.
- Jacobs, J. E., & Potenza, M. (1991). The use of judgment heuristics to make social and object decisions: A developmental perspective. *Child Development*, 62, 166–178. <http://dx.doi.org/10.2307/1130712>
- Jenny, M. A., Keller, N., & Gigerenzer, G. (2018). Assessing minimal medical statistical literacy using the Quick Risk Test: A prospective observational study in Germany. *British Medical Journal Open*, 8, e020847. <http://dx.doi.org/10.1136/bmjopen-2017-020847>
- Johnson, E. D., & Tubau, E. (2013). Words, numbers, & numeracy: Diminishing individual differences in Bayesian reasoning. *Learning and Individual Differences*, 28, 34–40. <http://dx.doi.org/10.1016/j.lindif.2013.09.004>
- Kahneman, D. (2011). *Thinking fast and slow*. London, UK: Allen Lane.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430–454. [http://dx.doi.org/10.1016/0010-0285\(72\)90016-3](http://dx.doi.org/10.1016/0010-0285(72)90016-3)
- Kleiter, G. D. (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. http://dx.doi.org/10.1007/978-1-4612-4308-3_27
- Laplace, P.-S. (1781). Mémoire sur les probabilités [Essay on Probabilities]. *Mémoires de l'Académie royale des sciences de Paris*, 227–332.
- Laplace, P.-S. (1951). *A philosophical essay on probabilities* (F. W. Truscott & F. L. Emory, Trans.). New York, NY: Dover. (Original work published 1814)
- Lewin, K. (1936). *Principles of topological psychology*. New York, NY: McGraw-Hill. <http://dx.doi.org/10.1037/10019-000>
- Lindsey, S., Hertwig, R., & Gigerenzer, G. (2003). Communicating statistical DNA evidence. *Jurimetrics*, 43, 147–163.
- Luchins, A. S. (1942). Mechanization in problem solving: The effect of Einstellung. *Psychological Monographs*, 52, 1–95. <http://dx.doi.org/10.1037/h0093502>
- Mandel, D. R. (2014). The psychology of Bayesian reasoning. *Frontiers in Psychology*, 5, 1144. <http://dx.doi.org/10.3389/fpsyg.2014.01144>
- McDowell, M., Galesic, M., & Gigerenzer, G. (2018). Natural frequencies do foster public understanding of medical tests: Comment on Pighin, Gonzales, Savadori, and Giroto (2016). *Medical Decision Making*, 38, 390–399. <http://dx.doi.org/10.1177/0272989X18754508>
- McDowell, M., & Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychological Bulletin*, 143, 1273–1312. <http://dx.doi.org/10.1037/bul0000126>
- McKenzie, C. R. M., & Nelson, J. D. (2003). What a speaker's choice of frame reveals: Reference points, frame selection, and framing effects. *Psychonomic Bulletin & Review*, 10, 596–602. <http://dx.doi.org/10.3758/BF03196520>
- McNamara, J. M., Green, R. F., & Olsson, O. (2006). Bayes' theorem and its applications in animal behaviour. *Oikos*, 112, 243–251. <http://dx.doi.org/10.1111/j.0030-1299.2006.14228.x>
- McNamara, J. M., & Houston, A. I. (1985). Optimal foraging and learning. *Journal of Theoretical Biology*, 117, 231–249. [https://doi.org/10.1016/S0022-5193\(85\)80219-8](https://doi.org/10.1016/S0022-5193(85)80219-8)
- Misuraca, R., Carmeci, F. A., Pravettoni, G., & Cardaci, M. (2009). Facilitating effect of natural frequencies: Size does not matter. *Perceptual and Motor Skills*, 108, 422–430. <http://dx.doi.org/10.2466/pms.108.2.422-430>
- National Health Service. (2017). *NHS breast screening: Helping you to decide* [Patient information brochure]. Retrieved from <https://www.gov.uk/government/publications/breast-screening-helping-women-decide>
- Pérez-Escudero, A., & de Polavieja, G. G. (2011). Collective animal behavior from Bayesian estimation and probability matching. *PLoS Computational Biology*, 7(11), e1002282. <http://dx.doi.org/10.1371/journal.pcbi.1002282>

- Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, 68, 29–46. <http://dx.doi.org/10.1037/h0024722>
- Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children*. New York, NY: Norton. (Original work published 1951)
- Pighin, S., Girotto, V., & Tentori, K. (2017). Children's quantitative Bayesian inferences from natural frequencies and number of chances. *Cognition*, 168, 164–175. <http://dx.doi.org/10.1016/j.cognition.2017.06.028>
- Real, L. A. (1991). Animal choice behavior and the evolution of cognitive architecture. *Science*, 253, 980–986. <http://dx.doi.org/10.1126/science.1887231>
- Real, L. A., & Caraco, T. (1986). Risk and foraging in stochastic environments: Theory and evidence. *Annual Review of Ecology and Systematics*, 17, 371–390. <http://dx.doi.org/10.1146/annurev.es.17.110186.002103>
- Rouanet, H. (1961). Études de décisions expérimentales et calcul de probabilités [Studies of experimental decision making and the probability calculus] *Colloques internationaux du centre national de la recherche scientifique* (pp. 33–43). Paris, France: Éditions du Centre National de la Recherche Scientifique.
- Rousselle, L., & Noël, M. P. (2007). Basic numerical skills in children with mathematics learning disabilities: A comparison of symbolic vs nonsymbolic number magnitude processing. *Cognition*, 102, 361–395. <http://dx.doi.org/10.1016/j.cognition.2006.01.005>
- Ruscio, J. (2003). Comparing Bayes theorem to frequency-based approaches to teaching Bayesian reasoning. *Teaching of Psychology*, 30, 325–328.
- Sarnecka, B. W., & Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition*, 108, 662–674. <http://dx.doi.org/10.1016/j.cognition.2008.05.007>
- Schooler, L. J., & Anderson, J. R. (1997). The role of process in the rational analysis of memory. *Cognitive Psychology*, 32, 219–250. <http://dx.doi.org/10.1006/cogp.1997.0652>
- Schulze, C., & Hertwig, R. (2018). *A description-experience gap in statistical intuitions: Of smart babies, stupid adults, intuitive statisticians, risk-savvy chimps, and cognitive illusions*. Manuscript submitted for publication.
- Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Mahwah, NJ: Erlbaum. <http://dx.doi.org/10.4324/9781410601247>
- Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, 130, 380–400. <http://dx.doi.org/10.1037/0096-3445.130.3.380>
- Sher, S., & McKenzie, C. R. M. (2006). Information leakage from logically equivalent frames. *Cognition*, 101, 467–494. <http://dx.doi.org/10.1016/j.cognition.2005.11.001>
- Siegler, R. S., & Braithwaite, D. W. (2017). Numerical development. *Annual Review of Psychology*, 68, 187–213. <http://dx.doi.org/10.1146/annurev-psych-010416-044101>
- Silver, E. A. (1986). Using conceptual and procedural knowledge. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 181–198). Hillsdale, NJ: Erlbaum.
- Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology*, 41, 1–19. <http://dx.doi.org/10.1146/annurev.ps.41.020190.000245>
- Skagerlund, K., & Träff, U. (2016). Number processing and heterogeneity of developmental dyscalculia: Subtypes with different cognitive profiles and deficits. *Journal of Learning Disabilities*, 49, 36–50. <http://dx.doi.org/10.1177/0022219414522707>
- Soares, N., & Patel, D. R. (2015). Dyscalculia. *International Journal of Child and Adolescent Health*, 8, 15–26.
- Staatsinstitut für Schulqualität und Bildungsforschung München. (2018). *LehrplanPlus Gymnasium Bayern* [Teaching curriculum for Bavarian Gymnasium]. Retrieved from <https://www.lehrplanplus.bayern.de/fachlehrplan/gymnasium/10/mathematik>
- Starkey, P., Spelke, E. S., & Gelman, R. (1990). Numerical abstraction by human infants. *Cognition*, 36, 97–127. [https://doi.org/10.1016/0010-0277\(90\)90001-Z](https://doi.org/10.1016/0010-0277(90)90001-Z)
- Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Belknap Press of Harvard University Press.
- Stigler, S. M. (1999). *Statistics on the table: The history of statistical concepts and methods*. Cambridge, MA: Harvard University Press.
- Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331, 1279–1285. <http://dx.doi.org/10.1126/science.1192788>
- Thaler, R. H., & Sunstein, C. R. (2003). Libertarian paternalism. *The American Economic Review*, 93, 175–179. <http://dx.doi.org/10.1257/000282803321947001>
- Tversky, A., & Kahneman, D. (1980). Causal schemata in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (Vol. 1, pp. 49–72). Hillsdale, NJ: Erlbaum.
- von Aster, M. G., & Shalev, R. S. (2007). Number development and developmental dyscalculia. *Developmental Medicine & Child Neurology*, 49, 868–873. <http://dx.doi.org/10.1111/j.1469-8749.2007.00868.x>
- von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. New York, NY: Cambridge University Press.
- Wassner, C. (2004). *Förderung Bayesianischen Denkens* [Teaching Bayesian thinking]. Berlin, Germany: Franzbecker.
- Wassner, C., Martignon, L., & Biehler, R. (2004). Bayesianisches Denken in der Schule [Bayesian thinking in school]. *Unterrichtswissenschaft*, 32, 58–96.
- Weber, P., Binder, K., & Krauss, S. (2018). Why can only 24% solve Bayesian reasoning problems in natural frequencies: Frequency phobia in spite of probability blindness. *Frontiers in Psychology*, 9, 1833. <http://dx.doi.org/10.3389/fpsyg.2018.01833>
- Weiss, R. H. (2006). *CFT 20-R: Grundintelligenztest Skala 2 – Revision* [Culture Fair Intelligence Test]. Göttingen, Germany: Hogrefe.
- Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, 358, 749–750. <http://dx.doi.org/10.1038/358749a0>
- Xu, F., & Garcia, V. (2009). Intuitive statistics by 8-month-old infants. *Cognition*, 112, 97–104. <http://dx.doi.org/10.1016/j.cognition.2009.04.006>
- Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, 74, B1–B11. [https://doi.org/10.1016/S0010-0277\(99\)00066-9](https://doi.org/10.1016/S0010-0277(99)00066-9)
- Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114, 245–272. <http://dx.doi.org/10.1037/0033295X.114.2.245>
- Zacks, R. T., & Hasher, L. (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Betsch (Eds.), *ETC. Frequency processing and cognition* (pp. 21–36). New York, NY: Oxford University Press. <http://dx.doi.org/10.1093/acprof:oso/9780198508632.003.0002>
- Zhu, L., & Gigerenzer, G. (2006). Children can solve Bayesian problems: The role of representation in mental computation. *Cognition*, 98, 287–308. <http://dx.doi.org/10.1016/j.cognition.2004.12.003>

Appendix

Problems Used in Studies 1, 2, and 3

There were a total of six problems; the magic wand problem is explained in the text and not repeated here. The others are listed here in the icon array condition. The natural frequency conditions are identical except that the icon array is not shown.

1. "Out of every 30 people living in a small village, five work as firemen, and the other 25 do not work as firemen. Four of the five firemen have a moustache. Also, 15 of the 25 men who do not work as firemen have a moustache."



"Imagine you meet a group of people from the village with moustaches. How many of them are firemen?" ____ out of ____

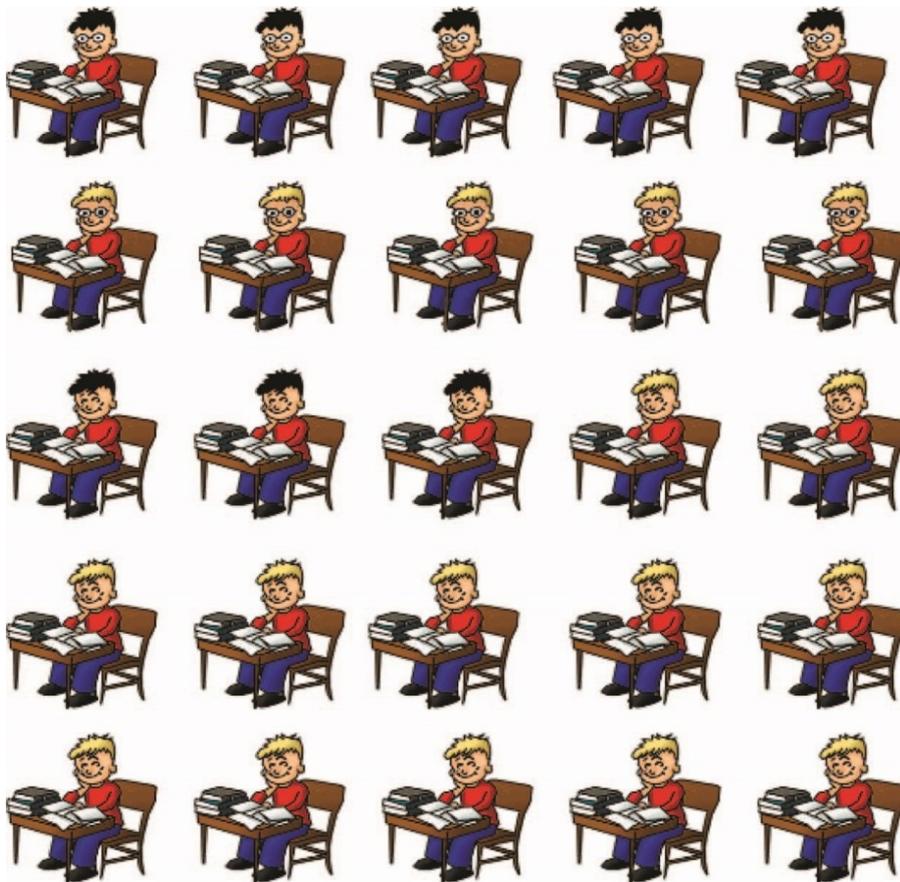
2. "Out of every 10 fairy creatures in a small and far-away fairyland, two are princesses, and eight are mermaids. Of the two princesses, one wears a crown. Also, two of the eight mermaids wear a crown."



"Imagine a group of fairy creatures from the fairyland who wear a crown. How many of them are princesses?" ____ out of ____

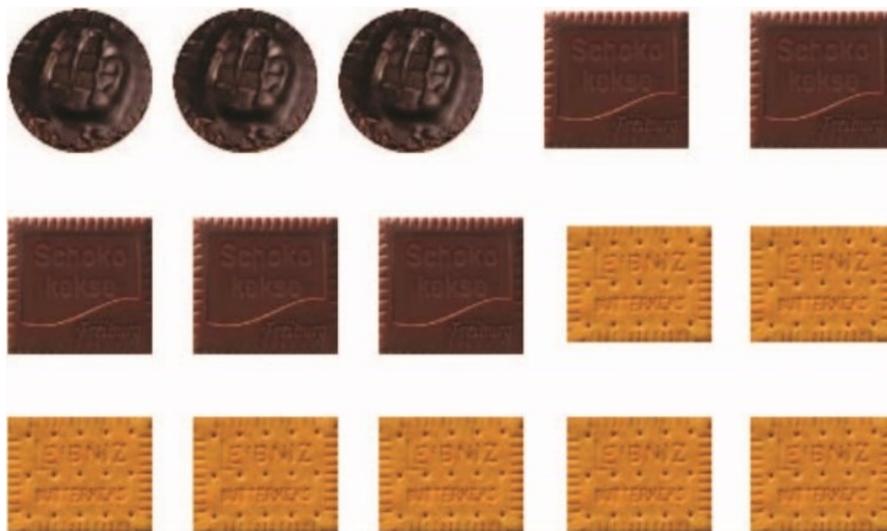
(Appendix continues)

3. "Of every 25 students in a class, 10 wear glasses, the other 15 do not. Of the 10 students who wear glasses, five have dark hair. Of the other 15 students, three have dark hair."



"Imagine a group of students from the class with dark hair. How many of them have glasses?" ____ out of ____

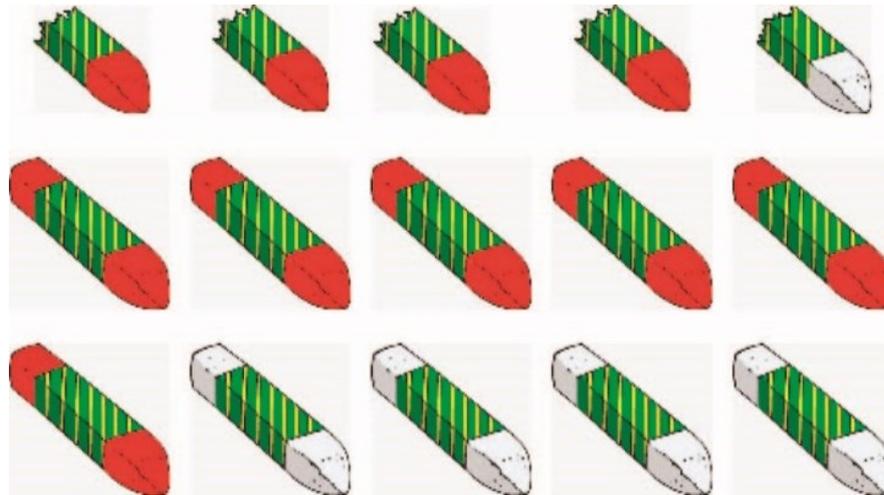
4. "Out of every 15 cookies in a cookie jar, three are round and made of chocolate. The other 12 cookies are square. Of the 12 square cookies, five are also made of chocolate."



"Imagine a pile of chocolate cookies from the jar. How many of these are round?" ____ out of ____

(Appendix continues)

5. "Out of every 15 chalk sticks in a box, five are broken and 10 are whole. Of the five broken sticks, four are red. Of the 10 whole chalk sticks, six are red."



"Imagine that you take out some red chalk sticks from the box. How many of them are broken?" ____ out of ____

Problems Used in Study 4

1. Thomas 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, eight have a red nose. Of the remaining 90 people who do not lie, 18 also have a red nose. Imagine that Thomas meets a group of people in the village with red noses. How many of these people will lie? ____ out of ____
2. 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." Seventy out of every 100 students understood. Of these 70 who understood, 63 put up their hands. Of the remaining 30 who didn't understand, nine put up their hands. Imagine a group of students who put up their hands. How many of them understood the principal? ____ out of ____
3. Twenty 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 do not 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? ____ out of ____
4. 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? ____ out of ____

(Appendix continues)

5. 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. Thirty 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, 36 of them are still red on the other side. Imagine you take out a group of red cards. How many of them have a cat picture on the other side? ___ out of ___
6. 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 headache in a hospital. How many of them get a cold? ___ out of ___
7. 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 who are not students of the university, three 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? ___ out of ___
8. 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? ___ out of ___