Please cite this text as: McDowell, M. & Jacobs, P. (in press). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychological Bulletin*. doi: 10.1037/bul0000126

Meta-Analysis of the Effect of Natural Frequencies on Bayesian Reasoning

Michelle McDowell and Perke Jacobs
Max Planck Institute for Human Development

September 6, 2017

Author Note

Michelle McDowell, Harding Center for Risk Literacy and Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development; Perke Jacobs, Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development.

We would like to thank Clara Schirren for her work and attention to detail when checking and coding studies. We would also like to thank Wolfgang Viechtbauer for excellent methodological advice, and Malte Lampart and Maia Salholz-Hillel for their work on the literature search and coding of studies. We thank Björn Meder, Charley Wu, and Gerd Gigerenzer for comments on prior versions of this manuscript.

Correspondence concerning this article should be addressed to: Michelle McDowell, Harding Center for Risk Literacy, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: mcdowell@mpib-berlin.mpg.de

Abstract

The natural frequency facilitation effect describes the finding that people are better able to solve descriptive Bayesian inference tasks when represented as joint frequencies obtained through natural sampling, known as natural frequencies, than as conditional probabilities. The present meta-analysis reviews 20 years of research seeking to address when, why, and for whom natural frequency formats are most effective. We review contributions from research associated with the two dominant theoretical perspectives, the ecological rationality framework and nested-sets theory, and test potential moderators of the effect. A systematic review of relevant literature yielded 35 papers representing 226 performance estimates. These estimates were statistically integrated using a bivariate mixed-effects model that yields summary estimates of average performances across the two formats and estimates of the effects of different study characteristics on performance. These study characteristics range from moderators representing individual characteristics (e.g., numeracy, expertise), to methodological differences (e.g., use of incentives, scoring criteria) and features of problem representation (e.g., short menu format, visual aid). Short menu formats (less computationally complex representations showing joint-events) and visual aids demonstrated some of the strongest moderation effects, improving performance for both conditional probability and natural frequency formats. A number of methodological factors (e.g., exposure to both problem formats) were also found to affect performance rates, emphasising the importance of a systematic approach. We suggest how research on Bayesian reasoning can be strengthened by broadening the definition of successful Bayesian reasoning to incorporate choice and process and by applying different research methodologies.

Public Significance

The present study shows that it is possible to improve people's inferences based on probabilistic information if conditional probabilities are presented as naturally sampled frequencies. Visual aids can help boost performance even further. However, future work is needed to understand why many people continue to have difficulties solving conditional probability problems.

The ability to make sound probabilistic inferences has long been considered essential to human rationality. Assessing whether human inferences adhere to the rules of probability theory has therefore a long tradition, not only in the study of judgement and decision making, but also in economics and philosophy (Gigerenzer et al., 1989; Hacking, 2006; Savage, 1954; von Neumann & Morgenstern, 1944). The conclusions from early work in psychology suggested that "man is an intuitive statistician", albeit a slightly conservative one (Peterson & Beach, 1967; Phillips & Edwards, 1966, p. 39). That is, studies suggested that human inference followed or approximated the rules of probability theory.

Subsequent work led to the contradictory view that the human mind was not built to work according to the rules of probability (Kahneman & Tversky, 1972, 1973). For example, based largely on laboratory studies using textbook problems, the heuristics-and-biases program documented a long list of cognitive errors or fallacies where human judgements about probabilities deviated from the normative standards of probability theory (for an overview see Gilovich, Griffin, & Kahneman, 2002). For example, one prominent finding was that participants tended to overweight or ignore base rates (e.g., the prevalence of breast cancer in a population) in probabilistic inference, phenomena referred to as the base-rate fallacy or base-rate neglect. In contrast, earlier work suggested that participants almost always take the base rate into consideration (see Koehler, 1996, for a discussion of theoretical and methodological criticisms of prior empirical work on the base-rate fallacy). Nevertheless, the apparent robustness of the base-rate fallacy was considered a demonstration of cognitive error (Bar-Hillel, 1980, 1984). This finding, among others, led to the conclusion that "[i]n his evaluation of evidence, man is apparently not a conservative Bayesian: he is not Bayesian at all" (Kahneman & Tversky, 1972, p. 450). These findings underlie the view of some behavioral economists that we can *nudge* people into making better decisions by exploiting their cognitive biases — a view often referred to as libertarian paternalism, as people retain choice but are steered towards decisions that governments or institutions deem welfare enhancing (Gigerenzer, 2015; Grüne-Yanoff & Hertwig, 2016; Thaler & Sunstein, 2008).

A number of explanations have been offered to reconcile these conflicting views. Some authors have argued that many findings from the heuristics-and-biases program use an inadequate normative standard, or that in many everyday situations it can be ecologically rational (e.g., adaptive in a natural ecology) to ignore information or contradict the axioms of rational choice theory (see, e.g., Gigerenzer, 1996a; Kahneman & Tversky, 1982; Stanovich & West, 2000).

Critiques have also compared the research methodologies used within different programs (Hertwig & Erev, 2009; Schulze & Hertwig, 2016) or have observed that prior studies have lacked ecological validity (Fiedler & von Sydow, 2015; Koehler, 1996). For instance, the apparent base-rate neglect phenomenon from research on textbook problems contradicts findings from related work on probabilistic reasoning using experience-based paradigms where participants learn the associations or co-occurrences between events. Animals, children, and adults in preliterate and prenumerate indigenous populations are found to be capable of making probabilistic inferences in line with the statistical properties of environments (Biernaskie, Walker, & Gegear, 2009; Fontanari, Gonzalez, Vallortigara, & Girotto, 2014; Gopnik, Sobel, Schulz, & Glymour, 2001; Rakoczy et al., 2014; Real, 1991; Real & Caraco, 1986; Sobel & Munro, 2009; Sobel, Tenenbaum, & Gopnik, 2004).

The present meta-analysis focuses on one of these criticisms that relates to the tendency for textbook tasks on Bayesian inference to ignore the connection between external information representations (e.g., numerical representations) and cognitive processing. In these textbook tasks, probabilities are summarised and one must make an inference based on a description of the relevant statistics. In their seminal paper, Gigerenzer and Hoffrage (1995) argued that there are different ways to summarise statistical information that are mathematically equivalent but not necessarily computationally equivalent, so that the choice of representation format affects the performance of a given cognitive process. For instance, although different numerical representations (e.g., Arabic or Roman numerals, or binary systems) can be mapped onto one another, the (cognitive) algorithms that operate on these representations may require different computations. A common analogy is that of the pocket calculator, designed to operate on Arabic numerals: can one infer that it has an algorithm for multiplication when it is fed information in binary numbers? In relation to an elementary Bayesian inference textbook task, often used in demonstrations of the base-rate fallacy, Gigerenzer and Hoffrage demonstrated that presenting statistical information in the form of joint frequencies resulting from natural samples, known as *natural frequencies* (defined in detail, below), yielded substantial improvements in Bayesian reasoning. The present meta-analysis focuses on research that has sought to account for or explain this facilitation effect.

Over the past 20 years, there has been some debate as to why natural frequencies can facilitate Bayesian inference in textbook problems. Although there is now a general consensus that natural frequencies can facilitate Bayesian inference (Brase & Hill, 2015; Johnson & Tubau,

2015), studies in this area report substantial variations in performance rates, ranging from 0 to 90 percent correct solutions recorded across studies. As such, it is unclear exactly how much natural frequencies facilitate performance and it is difficult to quantify the conditions under which facilitation effects are most likely to occur. For example, we know that natural frequencies can boost probabilistic inferences but to what extent and how high can performance be bolstered, given which features of the problem and in what contexts? More recent work has turned its attention to identifying the features of the person, problem, or methodological context that account for differences in performance rates across studies or that can shed light on underlying mechanisms. However, the field lacks a systematic examination of these factors. Rather, much of the prior work has focused on debating which theoretical perspective, the *ecological rationality framework* or *nested sets theory*, offers the most coherent account for why facilitation effects occur (Brase & Hill, 2015). Early work was plagued by misinterpretations of the natural frequency format, yet the theoretical accounts that were proposed on the basis of these misinterpretations actually converge on many common concepts (e.g., the importance of the subset problem structure). We move beyond these theoretical debates to examine why, when, and for whom natural frequencies facilitate Bayesian inference in an attempt to offer some closure or focus to the debate, and to provide guidance for future studies.

Accordingly, the current review and meta-analysis has three broad aims. First, we clarify what are natural frequencies (and what they are not) and highlight where theoretical accounts converge and diverge. Second, we identify and examine potential moderators for when, why, and for whom natural frequencies facilitate Bayesian inference. We report results from a meta-analysis of 35 studies representing 9611 participants on the relative effect of natural frequencies in comparison to conditional probabilities (normalised formats, defined below), and report how individual, methodological, and problem representation factors moderate this effect. Third, we emphasise current research gaps and suggest how to stimulate and progress research in this area.

Bayesian Inference and Natural Frequencies

Bayesian inference refers to the process of updating a prior probability of some hypothesis in response to new data. The normative benchmark for this process is provided by Bayes' theorem, described in detail below. A broad range of phenomena across many different research areas has been modeled using Bayes' theorem, from the study of perception to human cognition (Chater,

Oaksford, Hahn, & Heit, 2010; Chater, Tenenbaum, & Yuille, 2006). In relation to human cognition, the question is how well Bayes' theorem describes human inferences, for example, how the probability of some (unobserved) event changes in the presence of another (observed) event.

One approach to study Bayesian inference is the textbook paradigm where probabilities are summarised and one must make an inference based on a description of the relevant statistics. Consider, for example, an elementary textbook version of a Bayesian inference task presented using conditional probabilities (Eddy, 1982; Gigerenzer & Hoffrage, 1995, p. 685):

> The probability of breast cancer is 1% for a woman at age forty who participates in routine screening. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer? _____%.

In order to solve this task, let H denote the hypothesis (here, the presence of breast cancer), D a specific data outcome (here, a positive mammogram), and $\neg H$ and $\neg D$ their negations (here, the absence of breast cancer and a negative mammogram, respectively). The probability of the hypothesis after learning outcome D is given by

$$p(H|D) = \frac{p(D \cap H)}{p(D)} = \frac{p(D \cap H)}{p(D \cap H) + p(D \cap \neg H)}, \tag{1}$$

that is, by dividing the probability of having both breast cancer and a positive mammogram by the overall probability of having a positive mammogram (both with and without breast cancer). As the problem does not state the *joint probabilities* $p(D \cap H)$ and $p(D \cap \neg H)$, they need to be calculated. First, $p(D \cap H)$ can be obtained from multiplying $p(H)$, the base rate of breast cancer (that is, the prevalence of breast cancer in the reference class), with $p(D|H)$, the hit rate of the mammography test (the probability of a positive mammogram given that one has breast cancer). Similarly, $p(D \cap \neg H)$ is given by $p(\neg H) \times p(D|\neg H)$. Filling these probabilities into equation (1) yields Bayes' theorem:

$$p(H|D) = \frac{p(H) \times p(D|H)}{p(H) \times p(D|H) + p(\neg H) \times p(D|\neg H)}. \tag{2}$$

Generally speaking, Bayes' theorem describes how the *prior probability* $p(H)$, that is the probability of the hypothesis without additional information, is combined with a *likelihood* $p(D|H)$, that is the probability of outcome D if hypothesis H was true, to obtain the *posterior probability* $p(H|D)$, that is the probability of the hypothesis after obtaining additional information. Using the numbers given in the problem and equation (2), one can compute the solution:

$$p(H|D) = \frac{.01 \times .80}{.01 \times .80 + .99 \times .096}$$

$$\approx .078.$$

There is overwhelming evidence demonstrating that participants have difficulty solving such problems when presented in conditional probability formats, as shown above. Gigerenzer and Hoffrage (1995) reported that only 16 percent of participants in their study could provide the correct Bayesian solution to such problems, a finding consistent with many subsequent studies in the field (e.g., Chapman & Liu, 2009; Ferguson & Starmer, 2013; Mellers & McGraw, 1999). One interpretation of the poor performance on the above textbook problem is the base-rate neglect phenomena mentioned previously: participants neglect the base rate in their calculations and erroneously focus on specific case data (Barbey & Sloman, 2007).

An alternative view was offered by Gigerenzer and Hoffrage (1995). Drawing on an evolutionary perspective, they proposed that, as probabilities are a relatively recent information representation, probabilistic information would likely have been acquired and updated sequentially in reference to the event's frequency in natural environments — that is, a process of *natural sampling* (Kleiter, 1994). Gigerenzer and Hoffrage argued that cognitive algorithms for statistical inference would likely have evolved to operate on this type of representation. Consider, by analogy, a physician who acquires information sequentially about patients who have a symptom and a disease, and those who have the symptom but do not have the disease. The physician can then use these joint frequencies to calculate specific probabilities for a newly presented patient.

As we will see, the Bayesian algorithm is computationally simpler if probabilities are represented as using joint frequencies, which should allow more participants to find the correct solution. To test this assumption, Gigerenzer and Hoffrage (1995) presented participants with the same elementary Bayesian inference task described earlier but presented information in

natural frequencies (p. 688):

> 10 out of every 1,000 women at age forty who participate in routine screening have breast cancer. 8 of every 10 women with breast cancer will get a positive mammography. 95 out of every 990 women without breast cancer will also get a positive mammography. Here is a new representative sample of women at age forty who got a positive mammography in routine screening. How many of these women do you expect to actually have breast cancer? _____ out of _____.

Here, the *joint frequencies* are given in the problem and can be combined to obtain the posterior probability directly, analogously to equation (1),

$$\begin{aligned} p(H|D) &= \frac{n(D \cap H)}{n(D \cap H) + n(D \cap \neg H)} \\ &= \frac{8}{8 + 95} \\ &\approx .078, \end{aligned} \tag{3}$$

where $n(\cdot)$ denotes the frequencies. When these frequencies result from a natural sampling process, these are referred to as natural frequencies. Related work on experience-based probabilistic inference allows participants to experience this type of natural sampling process, whereas this version in the textbook paradigm presents participants only with the outcome of this process (Hoffrage, Krauss, Martignon, & Gigerenzer, 2015; Schulze & Hertwig, 2016).

As natural sampling preserves information about the base rate, which is contained in the joint frequencies, the base rate can be ignored, simplifying the calculation of the correct solution: stating the hit rate as 8 in every 10 women preserves the information that only 10 in every 1000 women have breast cancer; the information is not normalised[1] (e.g., compare Figures 1A and 1B to 1C and 1D, where the information is normalised in the former but not in the latter). When presented with information in the natural frequency format, Gigerenzer and Hoffrage (1995) found that 46 percent of participants were able to provide the correct Bayesian solution for these types of problems. Compared to the 16 percent who were able to solve the conditional probability problems in their study, this represents a considerable improvement, an effect we

[1]For consistency and clarity, in the present paper we use the term *conditional probabilities* to refer to any format that has a normalised structure, although we acknowledge that chances, percentages, and frequency formats with a normalised structure are not strictly conditional *probabilities*.

refer to as the *natural frequency facilitation effect.*

Clarifying the Natural Frequency Facilitation Effect

Following Gigerenzer and Hoffrage's (1995) study, a number of authors critiqued the notion that natural frequency formats facilitated Bayesian inference based on a misinterpretation of the format. Initial misconceptions related to the information structure, that is the distinction between naturally sampled frequencies, drawn from the concept of natural sampling by Kleiter (1994), and frequencies that are normalised or standardised. Natural sampling refers to the process by which one would naturally acquire information about events and their classes from experience, a sequential process where information is collated without fixing the marginal frequencies a priori (Gigerenzer & Hoffrage, 1995). In contrast, information formats that standardise or normalise information (see Figure 1A and 1B), do not preserve information about the base rates, which therefore require additional computation to incorporate them back into the calculation, as seen in Equation (2).

The natural sampling component was overlooked, rediscovered, and relabelled a number of times, often in order to incorporate the information structure within existing theoretical perspectives or to propose new theories (Gigerenzer & Hoffrage, 2007). For example, the natural sampling structure was introduced as the subset principle in the context of mental models theory (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999); the conjunctive information structure was rediscovered as a partitioned information structure[2] by Girotto and Gonzalez (2001) and Macchi (2000); and the observation that the numerator is a subset of the denominator was described as facilitating the construction of a set-inclusion mental model by Evans, Handley, Perham, Over, and Thompson (2000). This latter point was followed up by Sloman, Over, Slovak, and Stibel (2003) in support of a more general nested-sets hypothesis, or the argument that any representation that makes the nested (subset) relations transparent will facilitate performance. The nested-set observation is evident in Gigerenzer and Hoffrage's (1995) Equations 2 and 3, or the observation that the numerator is a subset of the denominator (Gigerenzer & Hoffrage, 1995, 2007).

It is now generally understood that information structure is central to the facilitation effect

[2]According to Girotto and Gonzalez (2001, p. 250), a partitioned structure is one where "the problem statement partitions a set of units into exhaustive subsets (e.g., a set of 100 people is partitioned into two subsets: 4 infected people versus 96 uninfected people; these in turn are divided into two subsets: 3 persons with a positive versus 1 person with a negative test result, and 12 persons with a positive versus 84 persons with a negative test result)."

(that is, the natural sampling structures in Figure 1C and 1D compared to the normalised structures of 1A and 1B). Nevertheless, nested-sets theory and the ecological rationality framework have been pitted against one another to debate which theoretical account provides the most homogenous explanation of findings to date. In the following section, we provide a brief overview of the ecological rationality framework and nested-sets theory to highlight points of distinction and to set the context for the different moderators of the natural frequency facilitation effect that have been proposed and that we review in our meta-analysis.

Theoretical Perspectives on Natural Frequencies

One aim of our meta-analysis was to review and quantify the predictions made by the ecological rationality framework and nested-sets theory in relation to the natural frequency facilitation effect. However, the many similarities in the predictions made by proponents of the two frameworks make it difficult to tease apart those that clearly differentiate the perspectives (see also, e.g., Brase and Hill, 2015, Johnson and Tubau, 2015, and McNair, 2015 for related observations). Admittedly, this problem was exacerbated by the fact that there is heterogeneity within the theoretical perspectives themselves and proponents differ as to the emphasis they place on certain concepts. In some cases, properties attributed to the theories have been imposed by others, often to the disagreement of the theory's proponents (see, e.g., comments on Barbey & Sloman, 2007). We review these theories with the general aim to highlight their contributions to the literature, summarise points of theoretical divergence, and to provide context for arguments made as to the relevance of different moderators. The meta-analysis ultimately moves beyond these theoretical debates and reviews the rich literature on the natural frequency facilitation effect to provide quantitative estimates and identify problem features or influential studies that can account for some of the variability in performance rates reported across studies.

Ecological Rationality Framework

The ecological rationality framework is a broad theoretical approach to the study of human cognition and decision-making (Gigerenzer, Todd, & the ABC Research Group, 1999) and has been applied to study of a range of topics including inference, choice, and group decision-making (Hertwig, Hoffrage, & the ABC Research Group, 2013; Todd & Brighton, 2015; Todd & Gigerenzer, 2007; Todd, Gigerenzer, & the ABC Research Group, 2012). Central to the framework is the concept of ecological rationality that emphasises the importance of consider-

ing the match between the human mind and the structure of the environment as a fundamental unit of analysis (Gigerenzer, 1998; Gigerenzer et al., 1999). On the one hand, this involves the study of cognitive mechanisms (e.g., the *adaptive toolbox* of decision strategies) and on the other hand it involves the study of the environmental structures that determine which of these mechanisms will be successful (*ecological rationality*; Todd & Gigerenzer, 2007). For example, the framework has been applied to study the role of recognition knowledge in inference, showing that simple recognition knowledge can promote accurate inferences specifically in environments where recognition is a valid cue (e.g., the recognition heuristic; Gigerenzer & Goldstein, 2011; Marewski & Schooler, 2011). With respect to Bayesian inference with textbook problems, the ecological rationality framework analysed different types of information representations (i.e., conditional probabilities and natural frequencies) from a computational perspective: The proponents showed that, although these informations formats are informationally equivalent, they do not entail computationally equivalent cognitive algorithms (i.e., Bayesian computations).

In the context of Bayesian inference, the ecological rationality framework has been attributed to Gigerenzer and Hoffrage (1995) and the study by Cosmides and Tooby (1996). Proponents converge on the idea that an evolutionary perspective is useful for identifying the match between cognitive strategies and environmental structures, offering a framework for investigating the adaptiveness of strategies to environments (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 2007). The proponents agree that external representations (e.g., probability information formats) influence internal representations (e.g., cognitive algorithms). However, proponents differ as to their views on the modularity of cognitive mechanisms. According to Cosmides and Tooby (1996; 2008), an independent, domain-specific cognitive mechanism evolved to operate on frequency representations and, when given the appropriate content (e.g., frequencies), this frequency module would "allow certain computations to proceed automatically or 'intuitively' and with enhanced efficiency over what a more general reasoning process could achieve given the same input" (Cosmides & Tooby, 2008, p. 66). As such, their perspective suggests an advantage for frequency representations more generally (e.g., also for normalised frequencies). In fact, in their widely cited paper on Bayesian facilitation often associated with the natural frequency facilitation effect, Cosmides and Tooby (1996) did not test natural frequency formats but rather variations on normalised frequency formats (e.g., formats similar in structure to 1B). Specifically, the authors investigated the strength of different numerical format manipulations (e.g., presenting a problem using normalised frequencies but asking for a response as a con-

ditional probability and vice versa) on Bayesian performance. Girotto and Gonzalez (2001) argued that the high performance achieved across studies in their paper is a consequence of the correct Bayesian response (2 percent) also potentially capturing participants who falsely divided the base rate by the total number of false positives, $1/50$ or $1/51$.

In contrast, claims about the modularity of cognitive mechanisms are generally not held in broader applications of the ecological rationality framework (Todd, Hertwig, & Hoffrage, 2005). Rather, Gigerenzer and Hoffrage's (1995) analysis of cognitive algorithms and information structures was computational and generated a priori theory-driven hypotheses about how cognitive processes map onto informational structures (the *ecological* aspect of the framework). In our view, the alignment between the theoretical perspective espoused by Gigerenzer and colleagues (Gigerenzer & Hoffrage, 2007) with the modularity view of Cosmides and Tooby (1996, 2008) has contributed to some of the critique of the ecological rationality framework from proponents of nested-sets theory.

Nested-Sets Theory

The most dominant alternative framework to account for the facilitative effect of natural frequencies on Bayesian inference is nested-sets theory. A general premise of the theory is that any representation that makes the nested-set structure of problems transparent will facilitate computations (Barbey & Sloman, 2007; Mandel, 2007). For example, one could use verbal description or visual representation to make the partitions and relations between relevant subsets clearer (e.g., a Venn diagram showing the relation between or nesting of subsets; Mandel, 2007). Nested-sets theory is founded in work on set relations in probability judgement and extensional reasoning. As such, its origins have been attributed to a variety of authors and theories (e.g., Barbey & Sloman, 2007; Girotto & Gonzalez, 2001; Johnson-Laird et al., 1999; Sloman et al., 2003; A. Tversky & Kahneman, 1983) and it claims to have broad applications to a variety of reasoning tasks (Barbey & Sloman, 2007; Mandel, 2007). Sloman and colleagues (2003) were the first to use the term *nested-sets hypothesis* and outline the theory in relation to probability problems. According to them, the most elementary relations are those related to set inclusion and set membership, and should these be made transparent in a representation, the arithmetic operations that follow are easier to perform. In this regard, proponents of the theory have made a number of predictions about alternative nested-set manipulations that can facilitate performance on Bayesian inference tasks, such as enhancing the transparency of nested

sets through visualisations (Yamagishi, 2003), and modifications to problem wording, such as improving the wording of text to show causal relations (Krynski & Tenenbaum, 2007; Sloman et al., 2003).

As nested-sets theory emerged, in part, in response to initial misconceptions about the natural sampling component of natural frequencies (outlined in the previous section), many of the current claims of the theory converge with the computational arguments initially put forward by Gigerenzer and Hoffrage (1995). In fact, at the time of Barbey and Sloman's (2007) review of research on base-rate neglect that incorporated work on natural frequencies, Mandel (2007) regarded nested-sets theory as "an assemblage of hypotheses, empirical findings, and rebuttals to theorists proposing some form of the 'frequentist mind' perspective" (p. 275). Proponents also disagree as to whether the theory should be situated within a dual-process framework (Barbey & Sloman, 2007; Evans & Elqayam, 2007; Lagnado & Shanks, 2007; Mandel, 2007; Samuels, 2007).

Nevertheless, despite recent calls for theory integration (Brase & Hill, 2015; Johnson & Tubau, 2015; McNair, 2015), nested-sets theory continues to be pitted against the ecological rationality framework on the basis of a few general claims. Specifically, in an effort to differentiate domain general reasoning processes from the strong modular domain-specific processes attributed to Cosmides and Tooby (1996, 2008) and the modularity principle, the theory has explored the effects of individual difference measures such as general intelligence (e.g., education, cognitive abilities) and motivation (e.g., incentivised performance; Barbey & Sloman, 2007; Lesage, Navarrete, & De Neys, 2013; Lesage et al., 2013; Sirota & Juanchich, 2011; Sirota, Juanchich, & Hagmayer, 2014). However, the finding that general intelligence or motivation is relevant to problem solving is not surprising to some authors who argue against the use of such data to infer cognitive structures or abilities (Brase, 2007; Brase & Hill, 2015; Trafimow, 2007). If evolutionary and dual-process claims were put aside, nested-sets theory and the ecological rationality framework appear to converge on the argument as to why natural frequencies support Bayesian inference: natural frequencies are a representation that provides a transparent information structure to simplify computations.

In summary, researchers from both nested-sets theory and the ecological rationality framework are interested in understanding the interaction or relation between external and internal representations of information (e.g., information formats and cognitive processes; Brase & Hill, 2015; Johnson & Tubau, 2015). Where divergence remains it relates to arguments as to how

these external representations influence internal ones (e.g., domain-general or specific cognitive mechanisms). A positive side of this debate has been the proposition of a variety of potential moderators to the natural frequency facilitation effect, which we discuss in detail below. The present meta-analysis quantifies the effect of these moderators with the aim to identify those features of the problem representation, methodology, and the individual that are most influential and can account for the substantial variability in performance across studies.

When, Why and for Whom is the Natural Frequency Facilitation Effect Most Likely to Occur?

Potential moderators of the effect are those that manipulate the problem representation (e.g., inclusion of visual aids), alter the methodological procedure (e.g., use of incentives), or account for aspects of the individual (e.g., expert or non-expert participants). Moderators related to problem representation can provide an understanding of the features of the format most influential to the effect. Methodological moderators are generally overlooked in debates about the facilitation effect, but nonetheless can contribute to the variation in the size of the effects reported across studies. Individual difference moderators can improve our understanding of the basic competencies of individuals that underlie effects and can help to tailor problem representations to different audiences. Previous reviews of this work have been qualitative, and although these have sought to provide a comprehensive overview of the literature to date, the conclusions of some of these reviews tend to support opposite theoretical viewpoints (Barbey & Sloman, 2007; Brase & Hill, 2015; Johnson & Tubau, 2015). A quantitative estimate of the magnitude of effects from different manipulations is currently lacking.

The current meta-analysis fills this gap, not only by isolating the effects of different moderators but also by providing quantitative estimates of the magnitude of effects. As such, the present meta-analysis focuses on those studies that have simultaneously compared performance for natural frequency and conditional probability formats. In the following sections, we review a broad range of potential moderators that have been put forth over the past two decades. In some cases, studies have examined the effect of moderators on the performance rates for one format only. We did not include these studies as they were more likely to differ in other ways that could not be controlled for in the analyses. While these studies cannot be included in the meta-analysis, we do mention these below as the moderators could still be coded for. We emphasise that not all moderators mentioned below were able to be included or coded for in

our meta-analysis and we specify why this is the case in the respective sections. Nevertheless, we think it is important to review all potentially relevant moderators to provide context for those that were included. Each potential moderator is indicated with italics within the following sections. Moderators that were able to be included in the meta-analysis are listed in Table 1.

Problem Representation

The most widely debated aspects of the natural frequency facilitation effect relate to the representation of problems, and the different ways of presenting information in natural frequency, conditional probability, or both formats. Over the years, researchers have modified formats by adding or substituting information, or by altering the computational complexity of the problems to tease apart potential explanations for the effect. While there are many aspects of the problem representation that have been explored, we attempt to address those features that have received most empirical attention.

Computational or problem complexity. The argument that natural frequency formats are less computationally demanding than conditional probability problems is generally not debated and in fact, this was the only argument made in Gigerenzer and Hoffrage's (1995) computational analysis. Specifically, to demonstrate that the computational advantage of natural frequencies lay in their more parsimonious segmentation of information, Gigerenzer and Hoffrage introduced the *short menu* of conditional probability and natural frequency formats. The short menu versions displayed only two pieces of information: $p(D)$ and $p(D \cap H)$ for the probability format and $n(D)$ and $n(D \cap H)$ for the natural frequency format. For example, the probability problem stated: "The probability that a woman at age forty will get a positive mammography in routine screening is 10.3%. The probability of breast cancer and a positive mammography is 0.8% for a woman at age forty who participates in routine screening". For the conditional probability version, the joint probabilities simplified the computational algorithm, $p(H|D) = p(D \cap H)/p(D)$, similar to Equation (1), while for the natural frequency version the computational algorithm was the same as Equation (3) above, albeit with the sum in the denominator already calculated[3]. As reported in the results of their Study 1, Gigerenzer and

[3]Mellers and McGraw (1999), Lesage et al. (2013), Fiedler, Brinkmann, Betsch, and Wild (2000) utilised a slightly different version of the short menu than Gigerenzer and Hoffrage (1995) which is known as the *joint menu*, where instead of providing $p(D)$, participants were provided with $p(H \cap D)$ and $p(\neg H \cap D)$ and they had to make the additional calculation: $p(D) = p(H \cap D) + p(\neg H \cap D)$ themselves. The natural frequency version is essentially the same, with information presented only in terms of joint frequencies with the complements provided.

Hoffrage found that when compared to the standard versions, the short menu versions resulted in improved performance for conditional probability formats, but not for natural frequency formats. Thus, the study showed that lowering computational complexity may explain the advantage of natural frequency formats over more computationally complex conditional probability formats. However, we note that in their study, short menu versions of natural frequency formats continued to outperform short menu versions of conditional probability formats despite similar computational algorithms: the natural frequency format improved from 46 percent for the standard format to 50 percent in the short format, and the conditional probability format improved from 16 percent to 28 percent correct solutions across problems.

While there is general agreement that short menu versions facilitate performance because there are fewer and easier calculation steps (Ferguson & Starmer, 2013; Mellers & McGraw, 1999), in particular for conditional probability formats, some authors have proposed alternative reasons for their advantage. For example, Mellers and McGraw (1999) suggest that presenting joint events makes the nested sets easier to visualise, thus facilitating computation. Fiedler et al. (2000) suggested that joint frequency and probability versions improved performance because these formats used a common reference scale (e.g., 8 *out of 1000* women; 95 *out of 1000* women), although we are unsure how the argument about the advantage of a common reference scale provides any additional explanation for this effect. Unfortunately, Fiedler et al. (2000) misinterpreted the natural frequency format used in Gigerenzer and Hoffrage (1995), suggesting that they had compared the short menu version to the standard conditional probability version and thus, this was responsible for the facilitation effect. Not surprisingly, Fiedler and colleagues found that short menu versions resulted in better performance than standard versions. Admittedly, it is difficult to tease apart theoretical arguments for the enhanced performance on short menu versions: making the nested set relations transparent is argued to require simpler computational algorithms, a prediction that was originally made by Gigerenzer and Hoffrage (1995). We distinguish between between short and standard menu versions in the coding of studies for the meta-analysis, to determine how much the facilitation effect is reduced as a result of the simpler computational structure.

To further address arguments regarding computational complexity, a number of studies have explored how manipulating the number and type of calculation steps affects performance. For example, Krauss, Martignon, and Hoffrage (1999) and Hoffrage, Krauss, et al. (2015) have argued that the information structure should still provide an advantage to natural frequencies

even as Bayesian computations become more complex with an increasing numbers of cues and hypotheses. That is, although overall performance would be expected to decrease with additional computational complexity, the natural frequency facilitation effect should still hold given multiple cue values (e.g., a medical test can return a positive, negative, or unclear result; in this case *three cue values*), multiple hypotheses (e.g., three diagnoses are considered; in this case *three hypotheses*), more than one cue (e.g., multiple tests, such as a mammography and ultrasound; in this case, *two or more cues*). Girotto and Gonzalez (2001) reported that creating an additional mathematical *calculation step* by providing $n(\neg H \cap \neg D)$ rather than $n(\neg H \cap D)$ for the natural frequency format, reduced performance rates (in their study, from 53 percent to 35 percent).

In this connection, Ayal and Beyth-Marom (2014) manipulated the number of calculation steps required to solve natural frequency problems by providing participants with different pieces of relevant or irrelevant numerical information and found that performance decreased as the number of steps increased (e.g., from 55 percent with one step to 10 percent with four steps). Unfortunately, similar manipulations for conditional probability formats are often not tested, and the relative detriment to performance cannot be compared. We intended to code for calculation steps in our meta-analysis, however, we found that it was not clear what would constitute a single mental step or if different mental steps should have equal weight (e.g., simple arithmetic versus recalculation or transformation). We do note that in one study minor manipulations to the numerical complexity of natural frequency versions by using large numbers or numbers not of multiples of 10 did not appear to reduce the facilitation effect (Misuraca, Carmeci, Pravettoni, & Cardaci, 2009). For our analysis, we code the number of cues and hypotheses within problems as an indicator of computational complexity.

A related manipulation has been to separate the problem format and question format such that the respondent is required to translate the information in the problem (e.g., natural frequencies) into a response for the alternative format (e.g., provide the answer as a probability). A number of studies have directly or unwittingly tested whether the facilitation effect is reduced if there is incongruence between problem and answer format (Ayal & Beyth-Marom, 2014; Evans et al., 2000; Fiedler et al., 2000; Girotto & Gonzalez, 2001). When the question asks for a numerical probability response, (i.e., *probability question*: What is the probability that a woman actually has breast cancer? _____%) this tends to reduce performance on natural frequency problems (Ayal & Beyth-Marom, 2014), yet when the question asks for a numerical response

using frequencies or pairs of integers (i.e., *frequency question*: How many of these women do you expect to actually have breast cancer? _____ out of _____), the effect on performance for probability problems is mixed (Cosmides & Tooby, 1996; Evans et al., 2000). Few studies directly test incongruent problem and question formats for both types of problems, instead they focus on frequency or probability problems only (Ayal & Beyth-Marom, 2014), or compare the effect of congruence between problem and question formats on probability and normalised frequency versions (e.g., Cosmides & Tooby, 1996; Evans et al., 2000). Nevertheless, the results of these studies suggest that a numerical response that does not match the information format of the problem reduces the natural frequency facilitation effect (Johnson & Tubau, 2015). Accordingly, it is expected that incongruence would disadvantage whichever format is made incongruent and thus, this may explain some of the variability in performance rates across studies. Where a direct comparison between natural frequency and conditional probability formats includes a manipulation of congruence for one or both formats, we include this moderator in the meta-analysis.

A number of other manipulations have been explored across studies, however in many cases the manipulations have been tested on a single format only or slight modifications to the formats have not been investigated systematically. Lesage et al. (2013) found that providing a total sample size or *enumerated population* on which to make calculations (e.g., "the study contains data from 8,500 children") did not enhance performance for probability versions, despite some indication that a reference class facilitates judgements on other Bayesian inference tasks (Neace, Michaud, Bolling, Deer, & Zecevic, 2008). Other studies state an enumerated population in probability versions but do not systematically test this format modification on performance rates (e.g., Fiedler et al., 2000; Konheim-Kalkstein, 2008; Macchi, 2000). However, meta-analyses can address this question by pooling estimates from different studies. It is unclear whether or not an enumerated population should enhance or decrease performance in the conditional probability format.

One final manipulation to problem complexity we think is worth mentioning involves textual modifications intended to increase or decrease problem comprehension. Increasing *verbal complexity* was shown to reduce performance (Johnson & Tubau, 2013), suggesting a role for basic text comprehension abilities in performance on word problems. Results of manipulations aimed at providing a *causal explanation* of a false alarm (e.g., that a benign but harmless cyst could cause a false-positive mammography result) to clarify the nested set structure or to facilitate

the construction of a causal model (and therefore the determination of relevant model parameters) to use in solving conditional probability problems are mixed (Krynski & Tenenbaum, 2007; McNair & Feeney, 2014, 2015; Sloman et al., 2003). Further, when Moro, Bodanza, and Freidin (2011) tested one of these causal explanation versions against a natural frequency format, they found no effect of wording on performance rates for this manipulation. Unfortunately, as few studies manipulated and tested performance across formats, the relative advantages of such manipulations from one format to another were not able to be examined systematically.

Multiple events. In earlier work on the natural frequency effect, the distinction between a conditional probability and single-event probability was often confused such that natural frequencies were thought to be an alternative to single-event probabilities (Barbey & Sloman, 2007; Cosmides & Tooby, 1996). However, conditional probability problems could be represented in terms of single events (e.g., probability between 0 and 1; chances of one event) or as *multiple events* (e.g., relative frequencies, see Figure 1A but with the phrasing "1% of women"), provided the information structure is normalised. To emphasise the different dimensions of statistical representations, Barton, Mousavi, and Stevens (2007) outlined a statistical taxonomy whereby representations can vary according to three orthogonal dimensions: number of events (single event or sets of events), numerical format (percentages, probabilities, fractions, or pairs of integers), and information structure (normalised or conjunctive). Thus, the statistical information in Bayesian inference tasks can be represented in multiple ways by crossing these dimensions.

To illustrate that the computational advantage of natural frequencies was related to the information structure rather than the number of events per se, Gigerenzer and Hoffrage (1995) demonstrated that the natural frequency facilitation effect did not extend to relative frequency versions of the conditional probability problems in their Study 2 (the relative frequency format tested by Gigerenzer & Hoffrage, 1995, was similar to Figure 1A but was presented in percentages and referred to women rather than a single woman). In further support of the relevance of information structure to the facilitation effect, Girotto and Gonzalez (2001) tested a representation that expressed chances of a single event, as pairs of integers (10 in 100 chances), in a format that mimicked the natural sampling structure of natural frequencies (compare C and D in Figure 1) and found that performance rates were similar for chance and frequency versions. However, simply using the term chance to represent problems does not seem to be sufficient to

improve performance. Brase (2008) tested natural frequencies against chance representations with a natural sampling or normalised structure to distinguish the effect of multiple events and information structure. Brase found that chances with natural sampling improved performance compared to normalised chances but that natural frequency formats were still the superior representation. People do appear to understand the distinction between probability and frequency (Girotto & Gonzalez, 2002), however, participants who interpret chances as frequency representations tend to perform better on these problems (Brase, 2008).

From a computational perspective, proponents of the ecological rationality framework argue that chances with natural sampling mimic the computations for natural frequencies in Equation (3) and thus, performance should be improved relative to normalised versions (Gigerenzer & Hoffrage, 2007; Hoffrage, Gigerenzer, Krauss, & Martignon, 2002). However, other authors disagree as to whether the adoption of the chance formulation as an extension of the natural frequency format fits within the ecological rationality framework and its evolutionary claims about the acquisition of information in the form of frequencies experienced as a series of events (Barbey & Sloman, 2007; Girotto & Gonzalez, 2002). Rather, proponents of nested-sets theory argue that, like natural frequency formats, chance formulations are effective because they make the set structure more transparent (Girotto & Gonzalez, 2001). Given the different ways in which information structure, numerical format, and number of events can differ in the problem representations used across studies, the current meta-analysis will code each of these three dimensions to discern those features of the representation most influential for facilitation effects. As the chances with natural sampling format is not normalised, it is fundamentally different from regular probability formats and accordingly, we do not examine these studies together with standard probability formats but conduct a separate analysis comparing them to natural frequency formats. Owing to the similarity in the natural frequencies and chances with natural sampling formats in terms of information structure and numerical format, it is expected that the performance is similar when comparing these formats, but we anticipate that there may still be an added advantage of frequencies, consistent with the work reviewed (see also Gigerenzer & Hoffrage, 2007; Hoffrage et al., 2002).

Calculations. As noted in the preceding sections, the formulation of the Bayesian inference problems can differ in a variety of ways and some authors have suggested that the specific numerical values for the *base rate*, *hit rate*, and *false-alarm rate* have the potential to influence

performance rates (see, e.g., Mellers and McGraw's, 1999, and Gigerenzer and Hoffrage's, 1999, discussion and reanalysis of the cab problem from Gigerenzer and Hoffrage, 1995). In a recent assessment of the influence of numeric task characteristics on solution rates, Hafenbrädl and Hoffrage (2015) examined how the base rates, hit rates, and false-alarm rates of 19 different Bayesian inference problems influenced performance rates using data from the 15 problems in Gigerenzer and Hoffrage (1995) and 4 problems from Hoffrage, Hafenbrädl, and Bouquet (2015). In general, they found that while higher base rates and hit rates were associated with a higher number of Bayesian solutions, higher false-alarm rates were associated with a lower number of correct responses. Further, probability formats were more affected by these factors.

In this connection, Gigerenzer and Hoffrage (1995) explored how different cognitive shortcuts could approximate Bayesian solutions under certain task conditions and proposed reasons for why higher base rates and hit rates are associated with better performance. For example, when the base rate, $p(H)$, is small (and $p(\neg H)$ is close to one), then the rare-event shortcut can be applied to simplify computation by approximating $p(D|\neg H) \times p(\neg H)$ by $p(D|\neg H)$. Few studies have explored alternative solution strategies or the environments in which they work well (Gigerenzer & Hoffrage, 2007). A few studies have attempted to classify the most common errors on Bayesian inference problems from supportive protocols (that is, written or verbal protocols accompanying solutions to word problems; Hoffrage & Gigerenzer, 1998; Siegrist & Keller, 2011; Zhu & Gigerenzer, 2006), however, few explore the connection between solution strategies and the specific features of the problems. As many studies report outcomes aggregated across problems, we cannot examine the specific effects of the base rate, hit rate, and false-alarm rate on performance, systematically. However, we believe the recent analysis by Hafenbrädl and Hoffrage (2015) sufficiently addresses these points.

Visual aids. Of the manipulations intended to bolster performance on Bayesian inference problems, the addition of a *visual aid* has been the most broadly tested (see Figure 2 for examples of the different types of aids tested across studies). Visual aids have been utilised to examine whether performance can be improved further by helping participants to visualise relations between the different pieces of information[4]. Specifically, some authors have argued that if one makes the nested-set structure transparent through visual aids (e.g., one can see how different events are related to one another), then this modification can reduce the natural

[4]Note that we use the term visual *aid* to denote that the visual is presented alongside or in addition to the text, not as a replacement.

frequency facilitation effect by enhancing performance for conditional probability formats (e.g., Sloman et al., 2003; Yamagishi, 2003). Others have argued that visual aids, shown to be effective in communicating probability information in other domains (e.g., health risks), can further enhance the effect of natural frequencies (Garcia-Retamero & Hoffrage, 2013). The relevance of visual aids for helping reasoners solve Bayesian inference problems was not only evident in the use of pictoral analogs by participants in Gigerenzer and Hoffrage (1995), but also in early work by Cole (1988) who explored different possibilities for visualising laboratory results to aid Bayesian reasoning. Generally, visual aids improve performance across formats (Brase, 2009a; Garcia-Retamero & Hoffrage, 2013; Sloman et al., 2003; Yamagishi, 2003). However, the theoretical arguments for when and why visual aids should or should not facilitate are mixed. Arguments attributed to the ecological rationality framework, typically the account by Cosmides and Tooby (1996), include that visual aids should facilitate performance when they incorporate discrete, individualised events that can be counted or when iconicity is high (i.e., greater similarity between the icon and the object or event it represents) because these help to tap into the frequency-encoding mechanism (Brase, 2009b, 2014; Sirota, Kostovičová, & Juanchich, 2014). Nested-sets theory proponents argue that it is the nested-set component of visualisations that facilitate performance, or that pictures can boost performance by drawing on people's visual computation abilities (Yamagishi, 2003).

To address some of these predictions, a number of studies have sought to illustrate that not all visual representations are equally effective (Brase, 2009b, 2014; Moro et al., 2011; Sirota, Kostovičová, & Juanchich, 2014). Different visualisation techniques have been employed, some of which may indicate a relative effect size indicated by the size of elements in the display (see, e.g., Figure 2B icon arrays or frequency grids and Figure 2D roulette wheel) and others that simply reveal set structure (see, e.g., Figure 2A frequency or probability trees). Moro et al. (2011) criticised the use of visual aids that confound clarification of set structure with the relative size of an effect. Testing two visual aid designs that did not reveal the relative size of an effect, (e.g., a Venn diagram that represented sets without reference to the relative size of the elements), Moro et al. (2011) found that these visual aids did not significantly improve performance for natural frequency or conditional probability formats. Rather, performance on natural frequency formats was consistently better than for conditional probability formats regardless of the use of visual aids. Further, as most participants in both natural frequency and conditional probability conditions were able to accurately identify the set structure of problems

illustrated with the visual aids, this suggests that the visual aids used in the study did not make the set structure of problems unclear. The present meta-analysis compares studies that use visual aids for both formats to those that do not use visual aids in order to establish whether visual aids reduce or enhance the natural frequency facilitation effect. We also code the specific type of visual aid employed.

Methodological Factors

Methodological or procedural factors may account for some of the variation across studies and a few studies have attempted to explore their effect on performance rates. For example, incentives to remunerate or motivate participants to complete a study (e.g., financial incentives, course credit) can enhance performance, particularly in cases where incentives are tied to performance (i.e., performance-based incentives; Cerasoli, Nicklin, & Ford, 2014). Brase (2009a) found performance-based incentives (*performance pay*) to be more effective than a *show-up fee* or course credit incentive structures at enhancing performance on Bayesian inference problems; there was no difference in performance for the latter two incentive structures. However, there are contradictory findings regarding the effect of incentives on performance according to presentation format. Sedlmeier and Gigerenzer (2001) found participants given rule-based training (Bayes rule) for solving conditional probability problems maintained high, stable performance at follow-up only when incentivised whereas participants receiving natural frequency training maintained performance regardless of incentives. Brase (2009a) found the effect of incentives was more evident for Bayesian inference problems of intermediate difficulty, notably variations of natural frequency formats, and not for the more difficult conditional probability formats. In contrast, Ferguson and Starmer (2013) found a main effect of incentives on performance but no interaction with presentation format. Given these mixed results and that theoretically it is not clear how incentives should influence performance across formats, we examine the incentive structure reported across all studies in the meta-analysis.

Multiple authors have suggested that the way in which Bayesian solutions are coded as correct may account for some of the variation across studies (Girotto & Gonzalez, 2001; McNair, 2015; McNair & Feeney, 2014). Studies vary as to whether a strict (point estimate) or more lenient estimate (point estimate plus or minus x percentage points) is used to classify correct Bayesian responses, and whether a supportive verbal or written protocol is required to determine Bayesian reasoning. Supportive protocols are used to classify marginally incorrect estimates that

suffer from minor calculation errors as correct, or to reclassify as incorrect those responses that indicate a correct guess but that have not followed a Bayesian reasoning approach. In Gigerenzer and Hoffrage's (1995) initial study, supportive protocols were used as an additional requirement for determining the use of Bayesian reasoning and to provide insight into the types of errors that participants made, thus distinguishing between outcome and process. We are aware of only one study that compared performance rates given different scoring criteria. McNair and Feeney (2014) found that a *strict scoring* (an exact estimate) resulted in near-negligible rates of correct responses when compared with scoring that allowed for calculation errors (within 5 percentage points of the correct estimate). The present meta-analysis attempts to address whether scoring criteria can account for some of the variation found across studies.

Given the wide variety of study designs and procedures used across studies on natural frequencies, we also consider other methodological factors that could account for variation in effects. For example, we consider the potential for learning or practice effects to influence performance rates and account for whether a *within-subjects* design was employed, whether participants solved a single or *both formats* (e.g., natural frequency and conditional probability formats), and record the total number of problems (*additional problems*) participants were required to solve.

Individual Differences

Characteristics of the sample or the individual have been investigated in an attempt to determine for whom the natural frequency format is most effective. Some of these individual-level moderators have been investigated with the view that they reveal insights into the nature of the cognitive mechanisms involved (e.g., general purpose or domain-specific; see Lesage et al., 2013; Sirota, Juanchich, & Hagmayer, 2014). Others have been explored in an attempt to elucidate boundary conditions or to identify characteristics of individuals that contribute to variations in performance (e.g., numeracy, education; see Brase, Fiddick, & Harries, 2006; Johnson & Tubau, 2013). For instance, exploring the natural frequency facilitation effect in different lay and expert populations can provide insights into how expertise, education, cognitive ability and development contribute to performance differences across formats and studies.

Individual differences have generally been explored in relation to differences in cognitive capacity or ability, for instance, in terms of educational achievement or cognitive development across the lifespan, to the importance of basic numerical skills or the role of expertise. There

is little surprise that higher general intelligence or cognitive ability (i.e., cognitive reflection, thinking disposition) has been linked with improved performance across formats for Bayesian inference tasks (Lesage et al., 2013; McNair, 2015; Sirota & Juanchich, 2011; Sirota, Juanchich, & Hagmayer, 2014). Nevertheless, even amongst individuals with higher ability there are substantial variations in performance across studies and considerable room for improvement. In this connection, numerous studies have explored whether a basic level of numeracy is needed for the facilitation effect to emerge (e.g., the threshold hypothesis; see Hill & Brase, 2012) or whether natural frequencies can facilitate performance even for individuals with low numerical ability (Galesic, Gigerenzer, & Straubinger, 2009). While high numerates tend to perform better across formats, results are mixed as to whether the effect of numeracy is independent to that of information format (Chapman & Liu, 2009; Hill & Brase, 2012; Johnson & Tubau, 2013).

Similar results are found in relation to the effect of educational experience on performance (Brase et al., 2006; Siegrist & Keller, 2011). For instance, even medical professionals and students, who are exposed to conditional probabilities in medical textbooks and curricula (e.g., test statistics, such as positive predictive value), have been shown to benefit from natural frequency formats (Friederichs, Ligges, & Weissenstein, 2014; Hoffrage & Gigerenzer, 1998), with some suggestion that professionals may benefit more than lay audiences (Bramwell, West, & Salmon, 2006). A number of studies have also shown that natural frequency formats facilitate performance on Bayesian inference tasks for both children (Zhu & Gigerenzer, 2006) and adolescents (Lesage et al., 2013), suggesting that the computational advantage is evident even at early stages of cognitive development. Galesic, Gigerenzer, and Straubinger (2009) also showed that natural frequencies facilitate performance in the elderly, who can face age-related cognitive declines in numerical reasoning abilities. Studies investigating training effects suggest performance on natural frequency formats is more robust over time (Kurzenhäuser & Hoffrage, 2002; Ruscio, 2003; Sedlmeier & Gigerenzer, 2001; Sirota, Kostovičová, & Vallée-Tourangeau, 2015a), indicating the potential to further build on cognitive ability with training.

In summary, general cognitive ability is associated with improvements in performance on Bayesian inference problems across a range of individual difference factors. However, it is unclear whether cognitive abilities increase the facilitation effect of natural frequencies or simply improve performance across tasks irrespective of format. Accordingly, we seek to quantify to what degree these individual characteristics contribute to performance differences across studies. Specifically, we code for a variety of individual differences across studies, including sample characteristics

(e.g., educational experience, or *expertise*) and whether the study made comparisons based on cognitive ability measures (e.g., *numeracy*). We anticipate that cognitive ability and educational expertise will increase performance rates for both formats and therefore it is unclear how these will influence the overall facilitation effect.

Overview of Meta-Analysis

As evident in our review, a variety of potential moderators have been proposed to account for the variation in the natural frequency facilitation effect found across studies. Some of these moderators have been introduced to explore or test specific theoretical accounts (e.g., cognitive ability, short menu format), while others have been proposed as factors to account for the variation in the size of effects across studies (e.g., scoring criteria). Predictions as to how each moderator should affect the size of the natural frequency facilitation effect are not always possible, because of mixed results and/or theoretical arguments in the literature (e.g., numeracy) or because studies have manipulated some moderators without testing them explicitly. Even in cases where moderators were not tested within any given study, meta-analyses are able to use the between-study variation to estimate the effect of such moderators, provided sufficient variation. Accordingly, the meta-analysis examines whether each moderator further enhances performance for natural frequencies, whether performance is enhanced across formats, or whether performance improves on conditional probability formats such that the relative difference in performance between formats is reduced. Table 1 summarises the moderators that were coded for in the present meta-analysis.

Method

Literature Search and Inclusion Criteria

Multiple methods were utilised to identify relevant papers for the review. A literature search of relevant databases (PsycINFO; PubMed; and Web of Science) was conducted on the following search terms (with thesaurus to explode terms, if available): bayes* AND conditional probab*, natural frequenc*, nested set, OR (reason* OR inference). Cited reference searches were conducted on key theoretical or review papers in the field (Barbey & Sloman, 2007; Cosmides & Tooby, 1996) as well as on Gigerenzer and Hoffrage's (1995) initial paper. Finally, a message was posted to the Society for Judgment and Decision Making (JDM) list requesting any addi-

tional publications, published or unpublished. No language restrictions were applied explicitly. The search covered papers published until December 2015, inclusive.

Studies were considered eligible for inclusion if they directly compared natural frequency and conditional probability formats. Studies that examined performance rates for different moderators on one format only were excluded. For studies that included multiple variations of different formats, the formats that were most equivalent to one another were selected for comparison (e.g., when both formats included a visual aid). If the moderator applied to one format only (e.g., the inclusion of an enumerated population for probability conditions), the effect was included and coded accordingly (see Table 1). Where more than one comparison was possible, for example when a natural frequency format was compared with two different conditional probability versions (typically normalised frequencies), the effect for the two most equivalent formats were compared, or two separate effects were included and then multiple comparison effects were controlled for in the analysis. We included studies that compared natural frequencies and chances with natural sampling, however these were analysed separately for the reasons specified in the review and again below (see section on *Coding of Moderators*; Brase, 2008; Girotto & Gonzalez, 2001; Sirota, Kostovičová, & Vallée-Tourangeau, 2015a). For the meta-analysis the sample sizes and performance rates for each format were required. When these were not reported in the publication, we attempted to obtain the relevant data from the corresponding author.

Screening for Eligible Studies

The database search yielded 213 relevant abstracts of which 35 reported on studies that met the inclusion criteria. The primary reasons for excluding studies were that they did not report an experiment (65 papers; 37% of excluded papers), studied a different type of probability problem (e.g., not conditional probabilities; 49 papers, 28%) or did not include one of the relevant formats for comparison (e.g., tested only natural frequency or probability problems; 44 papers, 25%). One additional paper was identified from the cited reference search, two additional papers were retrieved from the JDM mailing list (one currently unpublished), and one additional paper was sourced from the bibliography of another relevant paper during coding (the paper was published in a law journal and was not indexed for retrieval from the database search). For two of the relevant papers, an effect size could not be derived from the results or retrieved from the authors (Garcia-Retamero & Hoffrage, 2013; Kochetova-Kozloski, Messier,

& Eilifsen, 2011) and these could not be included in the meta-analysis. Three papers compared natural frequencies and chances with natural sampling formats (Brase, 2008; Girotto & Gonzalez, 2001; Sirota, Kostovičová, & Vallée-Tourangeau, 2015b). As two of these papers did not also include a conditional probability format as a comparison (Girotto & Gonzalez, 2001; Sirota, Kostovičová, & Vallée-Tourangeau, 2015b), these papers were coded separately and all chances with natural sampling conditions were analysed in a separate analysis (see subsection *"Subset analysis: Chances with natural sampling"*). Thus, a total of 35 papers that compared natural frequency and conditional probability formats were finally included in the meta-analysis.

Coding of Moderators

Each study was coded according to a coding manual that was developed on the basis of the literature reviewed. Binary variables were coded for the presence (1) versus absence (0) of the moderator. Basic study characteristics included year of publication (and publication status), sample size per group, sample characteristics (e.g., age and gender; these were not always available), and study design (within- or between-subject). Coding of papers was completed by research assistants and authors, with any disagreements resolved through discussion. One sixth of papers were double-coded and for each variable the agreement assessed using Cohen's κ, with a median value of 1 and an average of .89.

Problem representation. Problem representation moderators were recorded for each format separately (e.g., question format used in natural frequency format and in probability format), unless the moderator applied to only a single format (e.g., as in the case of an enumerated population). In studies where multiple problems were tested but the details for all problems were not provided, we made the assumption that all problems contained the same information, structure, and other representation factors (e.g., numerical format). Information structure or *short menu* was coded as either (0) for standard or (1) for short menu. A problem was coded as short menu if it presented only joint events, either with the denominator already calculated, as in Gigerenzer and Hoffrage's 1995 original study: $p(H \cap D)/p(D)$; or requiring a simple calculation as in Mellers and McGraw (1999): $p(H \cap D)/[p(H \cap D) + p(\neg H \cap D)]$. For the latter studies, the natural frequency version presented all joint frequencies (e.g., hit rate, false-alarm rate, and their complements) relative to a total sample rather than to subsets and the computation was the same as in standard versions. We coded studies that compared

both formats in short menu, or both formats in standard menu. Task complexity was coded for the number of hypotheses, cues, and cue values presented in the problem. We set out to code the following variations: (a) two hypotheses, single dichotomous cue; (b) two hypotheses, two dichotomous cues; (c) three hypotheses, single dichotomous cue; (d) two hypotheses, single cue with 3 cue values; and (e) two hypotheses, three dichotomous cues. As only one study (Hoffrage, Krauss, et al., 2015) examined a single problem for categories (d) and (e) (representing two effects each), and only two additional studies (Hill & Brase, 2015; Krauss et al., 1999) examined category (b) problems (representing 10 effects), these were combined under the more general *two or more cues* coding category, coded as present (1) or absent (0). This category represents more complex problem representations where two or more cues or cue values were examined (for example, multiple medical tests or a single test with multiple test results). *Three hypotheses* refers to case (c) where the problem included three hypotheses (e.g., one of three diagnoses) with a single dichotomous cue (e.g., a single medical test), and was also coded as present (1) or absent (0).

To account for any variation in effects attributable to the number of events referred to in probability formats, we coded moderator *multiple events* as (0) if the problem referred to the probability of a single event (e.g., the probability that a woman has breast cancer; chances that a student passes a test) and (1) if numbers referred to sets of events (e.g., the relative/absolute frequency of women having breast cancer). As all natural frequency problems relate to multiple events, this category only applied to the coding of the probability formats. We coded the numerical format used within each problem. However, there was minimal variation in the numerical formats used for the problems and questions. For conditional probability formats, only normalised frequency formats used pairs of integers (representing only 4 comparisons) whereas the remaining probability formats used percentages in the majority of cases, and in some cases fractions, or mixed numerical formats. Given the minimal variation in these numerical formats, we did not consider this moderator any further, except for cases in which an incongruent question format was used. In cases where a probability question was asked for the natural frequency format, we coded *probability question* as (1), and (0) otherwise. Similarly, when probability formats required a frequency response in pairs of integers, we coded *frequency question* as (1) and (0) otherwise.

An *enumerated population* (or reference class) was coded as present (1) if the probability problem included an enumerated population as part of the problem description, that is, a sample

size was provided for the participant to use in their calculations (e.g., "the data refers to 1,000 women"), otherwise it was coded as (0). For each estimate, the *base rate, hit rate*, and *false-alarm rate* used within the problem was recorded, where possible. If an estimate represented the average performance across multiple problems, or in cases where problems included multiple hypotheses or cues, no values were recorded. Studies that included a *visual aid* were coded for the type of visual aid used: (a) 2×2 contingency table, (b) frequency/probability tree, (c) Venn diagram, (d) picture, or (e) experienced-based interactive cards. Given that there were few instances of each type of visual, we compared effects that used any visual aid (a)–(e) coded as (1) versus no visual aid (0) in the analysis.

Three of the included studies (representing seven independent comparisons) made a direct comparison between natural frequency and chances with natural sampling problem formats. The only difference between the two conditions was related to the number of events (chances referred to the probability of a single event, and natural frequencies to multiple events). In all other respects, these chances formats differ almost entirely from other probability formats, specifically in terms of information structure (or lack of normalisation). We decided to include these effects when coding for studies because of recent interest in their similarities and the implications they are argued to have on theoretical perspectives (see, e.g., Brase, 2008; Sirota, Kostovičová, & Vallée-Tourangeau, 2015a). However, we analyse these effects separately from the primary analysis.

Methodological factors. For each effect, we coded sample size and whether the study employed a *within-subjects* design, coded as (1), or not, coded as (0), for use in analyses as well as various study procedural factors. We coded whether participants received at least one problem of *both formats* as (1), and (0) otherwise, and the number of *additional problems* completed per format. We also recorded whether problems were presented in a fixed sequence or random order, however most studies used counterbalanced designs and it was not possible to test any order effects. Further, we recorded whether an incentive was offered and if so, the type of incentive: A *show-up fee* was coded as (1) when a it was given in the form of money or course credit, and (0) if participation was voluntary. Similarly, *performance-pay* was coded as (1) when payment depended on performance, and (0) when participation was voluntary. Where no information about incentives was stated explicitly, no category was recorded.

In relation to *strict scoring* criteria, we categorised studies as applying either a strict or

more lenient criteria for scoring a respondent's solution as correct[5], and coded whether or not a supportive protocol was required. Specifically, we coded studies as requiring an: (a) exact estimate (include rounding), (b) exact estimate plus or minus a percentage margin, (c) exact estimate (including rounding) with supportive protocol, and (d) exact estimate (including rounding) or a supportive protocol[6]. Given that categories (a), (b), and (d) were determined to be more lenient criteria, and category (d) was not common, these were grouped together and *strict scoring* was coded (0) in these cases and (1) in case (c).

Individual differences. The primary sample of participants comprising each effect was coded to account for the education and expertise of the participants: (a) university undergraduate/postgraduate students, (b) medical students, (c) physicians, (d) general population, (e) older adults, (f) children, (g) secondary school students, (h) management executives. Owing to small samples across some of the categories, we considered categories (a), (b), (c), and (h) as instances of samples of (*experts*) in probabilistic reasoning or similar problems as described in the literature and coded these samples as (1) and (0) if they belonged to groups (d), (e), (f), or (g). Where *numeracy* was examined and reported as a moderator, samples were coded as having *high numeracy* (1) or low numeracy (0), typically determined by a median split of the sample in the study[7]. In all cases, the 11-item Lipkus numeracy scale (Lipkus, Samsa, & Rimer, 2001) was used to measure numeracy, although one study (Galesic, Gigerenzer, & Straubinger, 2009) added an additional item about a coin toss from Schwartz, Woloshin, Black, and Welch (1997). There were almost no studies that employed other measures of cognitive ability (see Lesage et al., 2013; Sirota, Juanchich, & Hagmayer, 2014, for exceptions).

Calculation of Outcome Measures

The focus of the meta-analysis is on the proportion of correct responses obtained for each of the two formats, that is, the natural frequency and the conditional probability formats. As mentioned above, we only considered studies that compared performance across formats eligible for inclusion. For this reason, each condition of each experiment yielded two proportions when

[5]Only one study used average errors as a dependent variable (Fiedler et al., 2000), for which we obtained data classifying responses as correct or incorrect.

[6]A supportive protocol could be used to confirm a correct estimate, or to recode an incorrect estimate if the correct process was followed but a calculation error was made. In all but two studies (Vallée-Tourangeau, Abadie, & Vallée-Tourangeau, 2015; Zhu & Gigerenzer, 2006), the protocol was used to confirm a correct estimate.

[7]In one case, we obtained raw data from the authors in order to calculate separate effects for high and low numeracy (Vallée-Tourangeau, Abadie, & Vallée-Tourangeau, 2016).

two formats were compared, or three proportions when three formats were compared. We will refer to them as a tuple and assume, for the sake of exposition, that each tuple consists of two proportions, one for each format. Each study can yield either one tuple or several when it includes multiple conditions. Likewise, each of the 35 articles could include one study or several.

For each tuple, we calculated the proportion correct for both formats, using

$$p_F = \frac{c_F}{c_F + i_F} \qquad p_P = \frac{c_P}{c_P + i_P} \tag{4}$$

where subscripts F and P denote the natural frequency format and the conditional probability format, respectively, c denotes the number of correct responses, and i denotes the number of incorrect responses. Each observed proportion provides one estimate of the true proportion underlying the observed value. We therefore refer to observed proportions as estimates.

The meta-analytic model that aggregates the different estimates requires that they follow normal distributions. However, because p_F and p_P can only take values in the limited range of $[0, 1]$, they cannot be normally distributed and require transformation before being aggregated. One common transformation is the logit-transformation, defined as

$$l_F = ln\left[\frac{p_F}{1 - p_F}\right] = ln\left[\frac{c_F}{i_F}\right] \qquad l_P = ln\left[\frac{p_P}{1 - p_P}\right] = ln\left[\frac{c_P}{i_P}\right]. \tag{5}$$

For each format, the logit is the natural logarithm of the odds of a correct response, which is in turn defined as the ratio of correct to incorrect responses. When correct and incorrect responses are equally frequent, the odds are one and the logit takes a value of zero. When a correct response is more frequent than an incorrect response, the odds exceed one and the logit is positive. Conversely, more incorrect than correct responses imply a negative logit. From now on, we will refer to the logit when we would usually refer to a proportion.

Because each logit is estimated from a finite sample of participants, we record the corresponding sampling variance as a measure of its precision. Following Woolf (1955), these variances were calculated using

$$v_{l_F} = \frac{1}{c_F} + \frac{1}{i_F} \qquad v_{l_P} = \frac{1}{c_P} + \frac{1}{i_P}. \tag{6}$$

Most estimates were calculated from independent samples so that we assumed the covariances between them to be zero. However, some estimates were based on the same set of par-

ticipants, which happened in either of three cases. First, in five studies, results were reported separately for each problem although all problems were solved by the same set of participants. In this case, we averaged proportions correct across problems, yielding only one estimate for each format. Second, in five experiments (yielding 18 logits), the same set of participants was used for both formats, yielding a within-subject estimate of the facilitation effect. Finally, five experiments (yielding 40 logits) shared participants across conditions. In both these cases, we estimated the covariance between the two logits based on the same set of participants using a method suggested by Stedman, Curtin, Elbourne, Kesselheim, and Brookhart (2011),

$$cov(l_a, l_b) = n \times \frac{n \times s - c_a \times c_b}{c_a \times c_b \times i_a \times i_b}, \tag{7}$$

where a and b denote two arbitrary formats, n denotes the sample size, and s denotes the number of participants giving a correct response under both formats[8].

We are not only interested in aggregating the proportion of correct responses for each format but also in the relative advantage of one format over the other. This effect size is captured by the odds ratio,

$$\text{OR} = \frac{c_F/i_F}{c_P/i_P}, \tag{8}$$

which gives the ratio of the odds for the natural frequency format to the odds for the conditional probability format. The logarithm of the odds ratio is easily computed from the logits of both formats,

$$ln[\text{OR}] = ln[c_F/i_F] - ln[c_P/i_P] = l_F - l_P. \tag{9}$$

Intuitively, $ln[\text{OR}]$ is positive when the logit (and therefore the proportion correct) under the natural frequency format exceeds the logit under the conditional probability format, and is negative in the reverse case.

After averaging proportions and calculating covariances, we obtained a final set of $k = 226$ logit estimates from 115 conditions in 35 papers, as well as a $k \times k$ covariance matrix of these logits, $\mathbf{V}$. The estimates could then be aggregated to obtain summary estimates and examine the effects of different study characteristics. The following section will present the statistical strategy for conducting this meta-analysis.

[8]There were eight logits for which the quantity s was not available. In three of these cases, either $c_F = 0$ or $c_P = 0$, so that $s = 0$. In the remaining five cases, we used the middlemost possible value and examined the effect of this choice on model outcomes (see section *Model Diagnostics*).

Aggregation of Outcome Measures

To aggregate the $k = 226$ logits shown in Table 3, we use a bivariate mixed effects model (van Houwelingen, Zwinderman, & Stijnen, 1993). Because bivariate models are relatively new to the meta-analytical toolbox, this section gives a detailed exposition of the model.

According to the bivariate model, each experiment is characterized by two logits that represent the proportions correct for each format. Rather than combining these logits in an effect-size estimate, the model treats them separately and estimates how a given moderator affects responses under each format. Formally, each experiment produces a tuple Y_j,

$$Y_j = \begin{pmatrix} l_{F,j} \\ l_{P,j} \end{pmatrix} = \begin{pmatrix} \theta_{F,j} \\ \theta_{P,j} \end{pmatrix} + \begin{pmatrix} e_{F,j} \\ e_{P,j} \end{pmatrix}, \tag{10}$$

which consists of two observed logits, l_F and l_P that serve as estimates of the true logits, θ_F and θ_P. Because estimates are based on random samples of participants, the true logits differ from their observed values by residuals e_F and e_P. The residuals are assumed to reflect only sampling variation and follow a multivariate normal distribution with covariance matrix

$$\mathbf{V} = \begin{pmatrix} v_{l_1} & cov(l_1, l_2) & \cdots & cov(l_1, l_k) \\ cov(l_2, l_1) & v_{l_2} & \cdots & cov(l_2, l_k) \\ \vdots & \vdots & \ddots & \vdots \\ cov(l_k, l_1) & cov(l_k, l_2) & \cdots & v_{l_k} \end{pmatrix} \tag{11}$$

where the (estimated) sampling variances of each logit are given along the main diagonal and the (estimated) covariances between the logits are either assumed to be zero or estimated using Equation (7), as discussed above.

The true logits of each estimate j are further assumed to consist of two components, reflecting the effects of moderators and residual heterogeneity. Specifically, we decompose the true logits into the following linear combination,

$$\begin{pmatrix} \theta_{F,j} \\ \theta_{P,j} \end{pmatrix} = \begin{pmatrix} \beta_{F,0} \\ \beta_{P,0} \end{pmatrix} + \begin{pmatrix} \beta_{F,1} & \beta_{F,2} & \cdots & \beta_{F,12} \\ \beta_{P,1} & \beta_{P,2} & \cdots & \beta_{P,12} \end{pmatrix} \times \begin{pmatrix} x_{1,j} \\ \vdots \\ x_{12,j} \end{pmatrix} + \begin{pmatrix} u_{F,j} \\ u_{P,j} \end{pmatrix}, \tag{12}$$

which describes each format's true logit as the sum of a base logit, the effects of moderators,

and the effect of residual heterogeneity. Consider first the base logits, $\beta_{F,0}$ and $\beta_{P,0}$. For each format, they give the average true logits underlying estimates from a standard study that has none of the 12 characteristics examined in the full-sample meta-analysis.

Consider next the effects of moderators. For each of the twelve study characteristics, $x_{m,j}$ is a binary variable that takes value 1 when estimate j is based on a study with characteristic m, and 0 otherwise. A value of 1 on $x_{j,m}$ adds the products $\beta_{F,m} \times 1$ and $\beta_{P,m} \times 1$ to the base logits, implying that $\beta_{F,m}$ and $\beta_{P,m}$ capture the effects of study characteristic m on the true logits. In contrast, a value of 0 on $x_{j,m}$ leaves the true logit unaffected by study characteristic m.

Finally, consider the effects of residual heterogeneity, $u_{F,j}$ and $u_{P,j}$. They capture, for example, differences in study design that are left unexplained by the set of moderators. Because studies differ in the exact problems they use and because these problems vary in difficulty, such residual heterogeneity is realistic. The mixed-effects model assumes that the effect of residual heterogeneity follows a bivariate normal distribution with covariance matrix

$$\mathbf{T} = \begin{pmatrix} \tau_F^2 & \tau_{FP} \\ \tau_{FP} & \tau_P^2 \end{pmatrix}, \tag{13}$$

where τ_F^2 denotes the variance of $u_{F,j}$, τ_P^2 denotes the variance of $u_{P,j}$, and τ_{FP} denotes their covariance. In addition, the model assumes that the sample of studies is a random sample from the population of possible studies. Provided that these assumptions are met, the mixed-effects model can be used to make inferences about the population of hypothetical studies. An alternative approach does not include random effects from residual heterogeneity. Such a fixed-effects approach does not allow inferences beyond the sample of studies included (Hedges & Vevea, 1998) and is therefore not considered for the present analysis.

The goal of the present meta-analysis is to estimate all β and τ coefficients. That is, we want to estimate both base logits, the average effects of all twelve study characteristics, and the (co-)variances of the random effects of residual heterogeneity. To assess the precision of these estimates, we compute cluster-robust standard errors that account for potential associations between results reported in the same paper (Hedges, Tipton, & Johnson, 2010). To estimate the parameters, individual logits are weighted by their precision using the covariance matrices $\mathbf{V}$ and $\mathbf{T}$. This implies that logits based on large, independent samples receive, ceteris paribus, higher weight than those based on small, correlated samples.

Results

In this section, we examine the pool of collected studies and report results of our meta-analytical model that aggregates the individual study results and delineates the effects of different study characteristics. This model was estimated using the statistical package R, version 3.3.3 (R Core Team, 2014), and the development version of the `metafor` package (Viechtbauer, 2010).

Distribution of Observed Values

Of the $k = 226$ logits collected for analysis, $k_F = 111$ concern the natural frequency format and $k_P = 115$ concern the conditional probability format. The disparity of k_F and k_P is due to the fact that in four cases an alternative probability format (normalised frequencies) was included along with the usual conditional probability format.

Figures 3 and 4 show forest plots of the proportions correct for the natural frequency and conditional probability formats, respectively. Here, each estimate is represented by a square, enclosed by its 95% confidence intervals. The distributions of observed proportions differ markedly between formats, with the majority falling above .2 in Figure 3 and below .2 in Figure 4. Proportions from the conditional probability format therefore appear to be smaller, on average, than those from the natural frequency format.

To examine evidence of selective reporting of studies, Figure 5 shows a funnel plot of the observed effects. Selective reporting refers to a biased publication system that favors results of a particular direction of the effect and/or size of the p-value. Because the effect size is given by the odds ratio, we estimated a univariate mixed-effects model (Hedges & Vevea, 1998) of the estimated ln[OR]. This model contained the same set of moderators as our original model but takes ln[OR] as the outcome variable. Figure 5 plots the residuals from this model against each estimate's standard error. One would expect estimates with small standard errors to have residuals closer to zero and those with large standard errors to fluctuate more strongly around the estimated value. The white area reflects this intuition and gives the 95%-confidence intervals.

Funnel plots are commonly used to detect asymmetries that indicate a relative lack of studies with residuals in a particular direction. Although the plot appears largely symmetrical, we can observe that the presence of two very imprecise studies with negative residuals in the lower

left portion of the plot creates a relative lack of studies with similar imprecision and positive residuals. Indeed, a formal test of funnel plot asymmetry in meta-analyses of odds ratios (Peters, Sutton, Jones, Abrams, & Rushton, 2006) rejects the null hypothesis of funnel plot symmetry at the one percent significance level ($t = 2.7$, $p = .0077$). Although such funnel-plot asymmetry cannot unambiguously establish selective reporting (see, e.g., Lau, Ioannidis, Terrin, Schmid, & Olkin, 2006), we conclude that there is some evidence of publication bias. Because over-reporting of studies with negative residuals has a negative effect on most of the estimated coefficients, we note that the estimates reported here may underestimate the underlying true effects. However, because the degree of asymmetry is small, we suspect that the under-estimation is limited[9].

Model Diagnostics

The meta-analytical model introduced in the previous section uses the observed proportions to estimate the underlying true proportions for studies with different combinations of study characteristics. Before we discuss the summary estimates, let us consider a few model diagnostics.

For each observed logit, the polygons in Figures 3 and 4 give the estimated average proportion for studies with the same characteristics, along with their 95% confidence intervals. However, the estimates do not include the effects of residual heterogeneity, that is heterogeneity in the true effect that is unaccounted for by the study characteristics included in the model. The model estimates $\hat{\tau}_F^2 = .54$, $\hat{\tau}_P^2 = .50$, and $\hat{\rho}_{FP} = .90$, where ρ_{FP} denotes the correlation between τ_F^2 and τ_P^2. The effect of residual heterogeneity is estimated to be, on average, larger in the natural frequency format than in the conditional probability format. At the same time, the high correlation between the random effects implies that studies with large residual heterogeneity in one format have, on average, large residual heterogeneity on both formats. A formal test of residual homogeneity ($Q_{188} = 773.19, p < .0001$) suggests that the presence of residual heterogeneity is not merely a sampling artifact. To put the estimated amount of residual heterogeneity into perspective, consider the ratio of residual heterogeneity to the total variation in observed logits (Higgins & Thompson, 2002; Jackson, White, & Riley, 2012). This ratio is estimated at $I_F^2 = 76\%$ for the natural frequency and $I_P^2 = 75\%$ for the conditional probability format. Thus, in both formats, more than 70 percent of the variance in effects is due to residual

[9]Ideally we would re-estimate the model after correcting for publication bias (e.g., Duval & Tweedie, 2000) or employ methods that are unaffected by publication bias (e.g., van Assen, van Aert, & Wicherts, 2015). However, these methods are either not suitable or not available for the bivariate mixed-effects model employed here.

heterogeneity, which would be considered a considerable amount of heterogeneity in medical research (Higgins, Thompson, Deeks, & Altman, 2003). Although psychological studies tend to be less stringently controlled, the amount of heterogeneity can nevertheless be considered large.

Like all model parameters, the amount of residual heterogeneity is estimated with some level of imprecision. We use the upper and lower bounds of the confidence intervals of all three heterogeneity parameters to examine the effect of this imprecision. In total, these parameter estimates can be combined in $2 \times 2 \times 2 = 8$ different ways. For each of these eight combinations, we re-estimate the model with τ_F^2, τ_P^2, and τ_{FP}^2 fixed a priori. Because the amount of heterogeneity affects the weights assigned to each observed logit, each re-estimation of the model yields a different set of estimated summary effects. However, none of the estimated changes in proportions varies by more than two percentage points, leading us to conclude that the imprecision in the estimated amount of heterogeneity has no considerable effect on the conclusions of our analysis.

Figures 3 and 4 show that accuracy appears to be lowest at the very top and bottom of each forest plot, where extreme observed logits are found. To identify such outliers, we computed the standardized residuals, dividing each residual by its standard deviation (Viechtbauer & Cheung, 2010). This metric follows a standard normal distribution where all logits with residuals $|z| > 1.96$ may be defined as outliers, 18 in the present case. Outliers can have an undue influence on the model parameters when their leverage is high, that is when their study characteristics differ strongly from the average. Leverage is expressed by an estimate's hat value and there are 4 logits with hat values indicating high leverage, however none of them is an outlier. Therefore, we would expect that none of the outliers has undue influence on model parameters. Indeed, we computed the Mahalanobis distances of the observed logits, which indicate their individual effects on the estimated summary values. Again, no single observed logit was found to have an undue effect on the model estimates, which may not be surprising given the total number of logits. Therefore, no observed logit was excluded from the meta-analysis.

Another sensitivity check concerns the estimation of the sample covariances in the section *Calculation of Outcome Measures*. As mentioned before, the estimation required knowledge of quantity s, the number of participants with a correct response in both formats. This number was not available for all studies but the range of possible numbers was restricted by zero on the lower end, and c_F and c_P (whichever is lower) on the upper end. From this range, we used the middlemost value for estimating the covariance. To examine the effect of this choice on the out-

comes of the meta-analysis, we re-estimated the model for all possible combinations of s values. Given the ranges of the five estimates, there are $2 \times 4 \times 10 \times 2 \times 6 = 960$ different possibilities and model re-estimations. The model parameters changed very little with no estimate of β_F or β_P varying by more than 0.1 and no value of τ_F^2 or τ_P^2 ranging by more than 0.02. We therefore suspect that this factor is unlikely to affect our conclusions.

Last, the model assumes normality in the errors and Figure 6 shows a quantile plot of the residuals, where the straight line indicates full normality. Indeed, the observed residuals follow the line closely and do not seem to deviate systematically, indicating an approximate normal distribution.

Summary Estimates and Study Characteristics

The upper panel of Table 2 shows the estimated summary effects. The first line gives the base proportions of both formats. Recall that the base proportions give the proportions correct for a standard study with none of the characteristics examined below. To obtain these proportions from the logits, we re-converted both estimates of β_0, using

$$\Delta_{F,0} = \frac{e^{\beta_{F,0}}}{1 + e^{\beta_{F,0}}} \qquad \Delta_{P,0} = \frac{e^{\beta_{P,0}}}{1 + e^{\beta_{P,0}}} \tag{14}$$

which is the inverse of Equation (5). For the natural frequency format, the model estimates that on average 24 percent of participants give a correct response. In contrast, on average only 4 percent of participants are estimated to give a correct response with the conditional probability format. These proportions imply an odds ratio of 7.1, which can be considered a strong effect. Even at the lower bound of the 95% confidence interval, this corresponds to an effect that is exceptionally large. On average, the share of participants who correctly solved Bayesian inference problems is 20 percentage points higher when they are presented in natural frequencies rather than conditional probabilities.

The remaining lines of the top panel in Table 2 give the changes in proportions that can be attributed to different moderators. For each format, we obtained Δ_m associated with study characteristic m by adding β_0 and β_m, converting the resulting logit into a proportion, and subtracting the base proportion,

$$\Delta_{F,m} = \frac{e^{\beta_{F,0}+\beta_{F,m}}}{1 + e^{\beta_{F,0}+\beta_{F,m}}} - \frac{e^{\beta_{F,0}}}{1 + e^{\beta_{F,0}}} \qquad \Delta_{P,m} = \frac{e^{\beta_{P,0}+\beta_{P,m}}}{1 + e^{\beta_{P,0}+\beta_{P,m}}} - \frac{e^{\beta_{P,0}}}{1 + e^{\beta_{P,0}}}. \tag{15}$$

Using the estimates of $\Delta_{F,m}$, $\Delta_{P,m}$ we are now able to examine the effects of each of the twelve study characteristics on the proportions correct and the resulting odds ratios. Note that a study characteristic that affects performance of both formats equally can have strong effects on the odds ratio, because it alters the relative advantage of one format over the other. For example, increasing performance from 50% to 60% in one format and from 10% to 20% in the other reduces the odds ratio from $\frac{.5}{.5}/\frac{.1}{.9} = 9$ to $\frac{.6}{.4}/\frac{.2}{.8} = 6$. Note also that we discuss the different study characteristics individually, although most studies include one or more of them simultaneously. For example, the original study by Gigerenzer and Hoffrage (1995) used *strict scoring*, tested *experts*, and had participants complete 14 *additional problems* in *both formats*. To obtain the proportions correct estimated by our model for this scenario, one needs to add the effects of all these moderators to the base proportions.

As shown in Table 2, *visual aids*, *three hypotheses*, and *short menu* were the strongest problem representation characteristics across formats. Methodological factors were also influential, in particular *both formats*. In many cases, moderators that improved performance for probability formats also improved performance for natural frequency formats. However, performance rates varied markedly across studies, and it is sobering to acknowledge that even with the inclusion of influential moderators, many participants were still unable to solve Bayesian inference problems in natural frequency formats. Below, we discuss the results and their practical and theoretical implications following our moderator analysis.

Problem representation moderators. Problem representation moderators affected natural frequency and conditional probability formats in a similar direction although to varying degrees and often with greater percentage increases for natural frequency formats. However, owing to the relative improvement or decline found for both formats, in some cases the size of the facilitation effect was reduced. A subset analysis on studies that compared natural frequencies to chances with natural sampling suggests that, despite a similar information structure, natural frequencies retained a facilitative effect.

Short menu. As predicted by Gigerenzer and Hoffrage (1995), short menu formats have a positive effect on probability formats but perhaps unexpected was the improvement in performance for the short menu versions of natural frequencies. A short menu is estimated to increase the share of correct responses by 12 percentage points in the natural frequency condition and by 11 percentage points in the conditional probability format. Although both proportions increase

by similar amounts, the parallel increase would lower the average odds ratio to 3.1 because the absolute difference in proportions now constitutes a smaller relative advantage. Thus, presenting short menu formats improves performance as predicted, but there is some added facilitative effect for short menu formats presented in joint frequencies.

As mentioned in the introduction, there were differences in the conjunctive structure of information presented in the short versions used across studies: three studies provided $p(D)$ and $p(D \cap H)$ (Ferguson & Starmer, 2013; Gigerenzer, 1996b; O'Brien, Roazzi, & da Graca B. B. Dias, 2004) and three studies provided $p(D \cap H)$ and $p(D \cap \neg H)$ (Fiedler et al., 2000; Lesage et al., 2013; Mellers & McGraw, 1999). Consistent with Gigerenzer and Hoffrage's (1995) argument, in both cases the equation for short and standard natural frequency versions remains the same (albeit in some cases with the denominator already calculated) whereas for conditional probability versions the equation is substantially simplified owing to the fact that the base rate is no longer needed in the calculation. Rather, participants must determine which of the joint events are relevant to the solution; this is most obvious in the case where only $p(D)$ and $p(D \cap H)$ are provided but may also be facilitated when all joint events are provided (Ottley et al., 2016; Wu, Meder, Nelson, & Filimon, 2016).

From a nested-sets theory perspective, Lesage et al. (2013) and Mellers and McGraw (1999) argue that formats presenting joint events facilitate performance because they help participants to visualise the nested or subset structure, although Mellers and McGraw (1999) argued that this would be the case for common but not rare events. It is not entirely clear how the results of the meta-analysis support or refute these claims, as the mechanism by which the subset structure is revealed has not been specified. Nor is it clear how the joint event formats help participants to visualise the nested structure. Similarly, the theory does not extend this explanation to further describe the improved performance of short natural frequency formats. Although the ecological rationality framework makes specific predictions for the facilitation of performance for short probability formats based on a computational analysis, the additional facilitation effect for natural frequency formats is also not immediately clear from a computational perspective. One potential explanation is that the base rate is not explicitly mentioned in the joint event or short menu versions which may reduce errors associated with selecting an incorrect denominator. To test this explanation would require primary data from these studies.

Three hypotheses. The improvement in performance across both conditional probability and natural frequency formats on problems involving three hypotheses was unexpected. While the natural frequency facilitation effect was anticipated to remain, the added complexity of the additional hypothesis was expected to decrease performance across formats. Introducing a third hypothesis increases performance in both formats, albeit to different extents. Performance for the natural frequency format is increased by 19 percentage points, and more than the 9 percentage point increase for the conditional probability format. Again, increased performance in both formats lowers the estimated odds ratio to 4.7, implying a reduced advantage of the natural frequency format over the probability format, but the effect remains considerable nonetheless. Looking at the three-hypotheses problems included in the meta-analysis gives some indication as to why this effect may have been found. Three papers examined three-hypotheses problems: Yamagishi (2003, contributing 12 logits) tested numerical variations of the gemstone problem, Johnson and Tubau (2013, contributing four logits) included a gemstone problem and an apple distribution problem based on the probabilities used in the gemstone problems in Yamagishi (2003), and Hoffrage, Krauss, et al. (2015, contributing two logits) tested a scenario where a medical test could identity the presence of one of two diseases (or neither disease). In all cases except for the Hoffrage, Krauss, et al. (2015) medical diagnosis problem (that represented two of the 18 logits), the problems segment information in a way that requires that only two hypotheses need to be considered for the solution. To illustrate, the gemstone problem segments information into three hypotheses (blurred, cracked, and clear gemstones) and the goal is to infer how many of the stones that pass inspection are clear. The hit rates and true negative rates are 100% for two of the hypotheses (a machine retains all clear stones and rejects all cracked stones), reducing it to a two hypotheses format with a perfect hit rate. For this reason Johnson and Tubau (2013) described the problems as simple because the format presents a simpler structure. This may explain the absence of a negative effect based on computational complexity predictions.

We can also further speculate as to why a positive effect was found for three-hypotheses problems. The numerical formats used in these problems, particularly for the conditional probability formats, differed from many of the other formats. In six of the three-hypotheses problems (representing 12 logits), common fractions were used to represent proportions (these were the only problems that ever used proper fraction representations), namely 1/2, 1/3, or 1/4. Calculations involving fractions are typically difficult mathematical operations to learn (Siegler,

Thompson, & Schneider, 2011), however, dividing quantities into common proportions or parts of a whole (e.g., half, third, quarter) is taught relatively early in formal schooling and fractions are frequently used to divide resources (e.g., sharing food). It is possible that the use of the part-to-whole representations afforded by the fractions and wording used within these problems facilitated the segmentation of information into proportions, similar to the effect of presenting joint frequencies (e.g., imagine cutting a cake in thirds, keeping a third, and cutting one of the thirds in half and comparing the two proportions). This interpretation is supported when considering the high performance on the conditional probability formats ($70-81$ percent correct responses) when the gemstone problems were also accompanied by the roulette wheel visual aid in the studies by Yamagishi (2003). The roulette wheel segments the information in such a way that its segments line up to create a clear visual segmentation, which may have facilitated performance in these conditions (Brase, 2014). In this connection, Yamagishi (2003) argued that such a visual aid could tap into people's visual computational abilities.

Two or more cues. Adding additional dichotomous cues or cue values (two or more cues) to the standard Bayesian inference task reduces performance for both conditional probability and natural frequency formats (by 4 and 2 percentage points, respectively), although the general facilitation effect of natural frequencies remains. Using two or more cues instead of one was estimated to have a small, negative effect on responses in the natural frequency format and a negligible effect on responses in the conditional probability format. Jointly, these two effects increased the implied odds ratio to 11.4. However, both effects are imprecisely estimated from the small set of studies that are currently available, and the results may change, in both sign and magnitude, with more diverse examinations.

Multiple events. In the present meta-analysis, studies that employed conditional probability formats with multiple-event phrasing had a small improvement on performance rates by 2 percentage points compared to problems with single-event phrasing[10]. Curiously, the framing of the conditional probability format appears to have affected performance in the natural frequency format, although these effects are imprecisely indicated and may reflect sampling variation. Taken together, the two effects decrease the estimated average odds ratio to 5.5. One explanation for these effects is that the types of problems that were used in these studies were

[10]To ensure the effect was not driven by the few studies that examined normalised frequency formats presented using numerical frequencies (e.g., 10 in 1000) we repeated the analysis with these studies excluded. The effect did not change.

generally easier problems. However, as we cannot examine the problems in greater detail, this is only speculative.

Incongruent question formats and enumerated population. Incongruent question formats (*probability question, frequency question*) appear to offer only small disadvantages to the respective formats (each decreases 2 percentage points). As these estimates are based on a small number of studies that manipulated these characteristics, the effects are estimated imprecisely and may require further examination. However, based on the evidence currently available, the effects appear to be small and negligible. Similarly, augmenting the probability formats with an *enumerated population* does not appear to facilitate performance (no percentage change). This effect is imprecise for natural frequency formats but is estimated with greater precision for the probability formats. Thus, small changes to the complexity of the problem or introducing a requirement to convert problem and question formats appear to have a negligible effect on performance, particularly in contrast to the strong effects of information structure.

Visual aids. The strongest moderator of performance for both natural frequency and probability formats involved the inclusion of a visual aid. Supplementing both formats with a visual aid increased performance by 23 and 22 percentage points for the natural frequency and probability formats, respectively. The strong improvements in performance for both formats decreased the odds ratio to 2.5, although this still indicates a strong natural frequency facilitation effect. When visual aids are used for conditional probability formats they enhance performance to a similar level as natural frequency formats without visual aids. However, given that visual aids improve performance on natural frequency formats to the same degree, it appears that they may have an independent effect to that of format. Visual tools have been used throughout history to convey meaning and represent relations between concrete and abstract concepts, are beneficial for understanding concepts in mathematics and problem-solving more broadly, and tend to be spontaneously produced when individuals attempt to solve probability problems (B. Tversky, 2001; Zahner & Corter, 2010). In this connection, visual aids have been shown to improve comprehension of health risks in risk communications (Galesic, Garcia-Retamero, & Gigerenzer, 2009; Garcia-Retamero & Cokely, 2013).

There have been several recent studies that have explored the features of visual designs related to performance on natural frequency formats. Micallef, Dragicevic, and Fekete (2012) explored different Euler diagrams and frequency grids that varied whether or not the area of

the different regions was proportional to the quantities represented in the problem. They tested six different variations and found that none of the variations were superior to one another, and only slightly improved performance rates compared to a natural frequency text. However, performance rates across formats in this study were generally poor. Khan, Breslav, Glueck, and Hornbaek (2015) investigated the different qualities of visualisations that could facilitate performance on Bayesian inference problems, again only with respect to natural frequency formats. Khan et al. (2015) distinguished between visualisations that emphasised branching or information structure (e.g., trees), nested-set relations (e.g., Euler diagrams), and frequencies or the quantities involved (e.g., icon arrays), and developed two hybrid diagrams that combined branching with frequency (a Sankey diagram) or branching that illustrated all sets (a double-tree). The frequency grid and double-tree diagrams resulted in the best performance (20 percent of participants answered correctly in both conditions), however performance was similar across conditions. Unfortunately, owing to the small samples of studies exploring visual aids, we were not able to explore the specific features of the visual aids that were more or less likely to facilitate performance in the present meta-analysis.

Subset analysis: Chances with natural sampling. Studies investigating the facilitative effect of chances with natural sampling also suggest that the interpretation of chances as frequencies rather than as a single event improves solution rates (Brase, 2008, 2014). In the present study, natural frequency representations demonstrate superior performance to chances with natural sampling. The second panel of Table 2 shows the results of a subset analysis of $k_{chs} = 18$ logits from studies examining chances. For this subset analysis, we used the same model as for the full-sample meta-analysis but excluded any study characteristics. The lack of study characteristics as covariates means that the model does not account for large parts of . As a consequence, the estimated amounts of residual heterogeneity are higher than in the original model ($\tau^2_F = .6, \tau^2_P = .7$) and the null hypothesis of residual homogeneity is rejected ($Q = 78.1, p < .0001$). The average performances are estimated at 40 percent and 19 percent, for the natural frequency and conditional probability format, respectively. The implied odds ratio is estimated at 2.8. This analysis is based on few observed logits and does not control for other study characteristics, so that the estimated proportions should not be compared with the results of the full-sample meta-analysis. Nonetheless, the results suggests that although the information structure is the same, single-event representations may still be a more difficult

representation for participants to solve.

One additional aspect to consider when interpreting this analysis is the difference in question format used within these studies. Girotto and Gonzalez (2001) argued that the question format for natural frequency versions prompts participants to compute two terms of the ratio (e.g., those who test positive, and those that test positive and have breast cancer) whereas probability versions do not, thus making the solution easier for natural frequency versions (in addition to information structure). For natural frequency and chances with natural sampling formats, two step questions that first require participants to provide the denominator of the ratio (e.g., people who test positive) followed by the numerator (e.g., people have breast cancer and test positive) were better than the standard question format (i.e., _____ out of _____) for both natural frequency and chances with natural sampling formats. In all of the studies comparing chances with natural sampling and natural frequency formats the two part question was used and may explain the higher performance for both formats in these studies (Brase, 2008; Girotto & Gonzalez, 2001; Sirota, Kostovičová, & Vallée-Tourangeau, 2015b). Only Brase (2008) used a two part question form when comparing normalised chances and natural frequency formats, limiting our ability to test the moderator in the present meta-analysis.

Methodological factors. Methodological factors were found to account for differences in performance for the different formats, suggesting that some of the variation in the natural frequency facilitation effect reported across studies may result, in part, from differences in study design. As we will see, some of the methodological moderators do not affect both formats in the same way, thus distorting the facilitation effect relative to a typical experimental setup. The results can therefore guide methodological decisions in future empirical work.

Both formats. Study designs that involved participants receiving one or more conditional probability problems *and* one or more natural frequency problems appeared to help performance for natural frequency but hinder performance for conditional probability formats. Exposing participants to both formats (e.g., in a within-subject design) is estimated to increase performance for natural frequency formats by 13 percentage points on average, and slightly decrease performance for conditional probability formats; this combination increases the odds ratio to 16.2, which is substantial. One potential explanation for these effects is that solving both types of formats may interfere with the strategies applied to subsequent problems. In some cases, this may have an advantage whereas in other cases it may not. As the majority of studies that used

multiple format designs also counterbalanced the order of problems, we are unable to test any order effects, and no order effects were reported in these studies.

Additional problems. Performance on natural frequency and conditional probability formats improved with the number of problems that participants were required to solve. The experimental set-up expressed in the baseline proportions assumed that each participant was given only one problem. Assuming linearity, the model estimates that each additional problem increases performance, on average, by around 1 percentage point for natural frequency format and .5 percentage points for the conditional probability format. Because some studies have their participants solve ten Bayesian problems, implying an increase in performance by 10 and 5 percentage points, respectively, the estimated effect due to practice can be considerable. In such cases, the odds ratio would decrease to 5.3. It may be that further opportunities to solve problems may increase the potential for one to try out different solution strategies that turn out to be successful for at least one of the problems or that the underlying structure of problems is more likely to be recognised. However, our hypotheses about the reason for this effect is purely speculative; rather we stress that methodological differences can account for some of the variation across studies.

Strict scoring criteria. Contrary to the results of McNair and Feeney (2014), there was no effect of a stricter scoring criteria on performance rates, as defined in the current study. McNair and Feeney (2014) compared scoring that involved either an exact estimate or an exact estimate $+/-5$ percentage points for conditional probability problems. In the reviewed studies, the $+/-$ % range varied across studies from $1-5$ percentage points (Bramwell et al., 2006; Chapman & Liu, 2009) making it difficult to determine an appropriate cut-off value to indicate greater or lesser leniency. Nevertheless, in the present study we sought to compare lenient against the strictest scoring criteria which we determined to be one that required not only a correct estimate but also a correct protocol to support the solution. The justification for these scoring criteria, as stated by the authors of the reviewed studies, was to ensure that a correct estimate was *not* a result of a guess or an alternative, non-Bayesian strategy. However, the strictness of the scoring criteria had a negligible effect on performance in both formats, with a decrease of 2 and a slight increase of 1 percentage points for natural frequency and probability formats, respectively. These estimates, particularly for the natural frequency format, are imprecise and may change when additional studies that employ strict coding criteria are

included.

Subset analysis: Incentives. Of the $k = 226$ logits in the full sample, data on incentives is available for only $k_{inc} = 165$ logits, of which 115 use show-up fees, 12 use performance pay, and 38 use neither. For this reason, we could not include incentives in the full-sample meta-analysis but examined them in a separate analysis that does not include other study characteristics as covariates. The model used for aggregating the logits is the same as before with incentive included as the only covariate. Again, the amounts of residual heterogeneity are higher than in the original model ($\tau_F^2 = 1.0, \tau_P^2 = 1.9$) and the null hypothesis of residual homogeneity is rejected ($Q = 1373.5, p < .0001$). The third panel of Table 2 shows the effects of incentives. The proportions for experiments without incentives are estimated at 41 percent for the natural frequency format and 10 percent for the conditional probability format. Unlike a show-up fee, which appears to have a negligible effect, performance pay is estimated to increase performance by 23 and 11 percentage points, respectively, again without controlling for differences in study designs, which complicates comparisons with the full-sample meta-analysis.

There were only two studies that systematically examined performance-based incentives (Brase, 2009a; Ferguson & Starmer, 2013, representing a total of 12 logits), with many studies incentivising participants with the award of course-credit or payment of a show-up fee. As stated previously, the two studies that investigated the effect of performance-based incentives on performance across formats found contradictory effects: Brase (2009a) found incentives facilitated performance on natural frequency problems whereas Johnson and Tubau (2013) found a general effect of incentives irrespective of format. Unfortunately, owing to the limited studies on performance-based incentives we cannot resolve this conflict. Rather, we can conclude that, similar to studies in other domains, performance-based incentives may generally improve performance across formats (Cerasoli et al., 2014).

Individual characteristics. Owing to only a few studies examining numeracy, we could only include experts as a moderator in the full-sample meta-analysis. However, we conducted a subset analysis on studies that included effects for low and high numerate participants.

Experts. Natural frequencies benefit experts and non-experts alike, and the results of the meta-analysis do not suggest that participants with greater educational or professional experience are better able to solve problems presented in either format, as compared to lay samples.

The slight increases in performance in both formats are imprecisely estimated and may change direction as further studies comparing expert and novice samples accumulate. However, we acknowledge that a limitation to our analysis is the reduction of a broad range of samples to a dichotomy of samples that were presumed to be experts and those that were presumed to be non-experts, although, the distinctions were consistent with arguments made within the literatures reviewed. Medical professionals, management executives and university or postgraduate students are often described as comprising more expert, educated samples. While we found that a broad range of samples were used across studies, the majority were based on university students, consistent with the general criticism of research in the behavioural sciences that studies tend to be based almost entirely on university student samples (Henrich, Heine, & Norenzayan, 2010). Expanding research to include more diverse samples could provide valuable insights into how Bayesian reasoning is learned and evolves over time. In particular, we think that a greater focus on Bayesian reasoning in children is warranted in order to explore developmental trajectories, for example, to elucidate which mathematical concepts are influential in revealing solution strategies (see, e.g., Zhu & Gigerenzer, 2006) or to learn where difficulties first emerge and how they can be targeted. In this connection, further work on older adults, who may suffer from cognitive declines with age, would be important from a more applied perspective, particularly given that older adults are increasingly faced with the implications of medical test results.

Subset Analysis: Numeracy. Of the $k = 226$ logits in the full sample, data on numeracy is available for only $k_{num} = 52$ logits. For this reason, we could not include numeracy in the full-sample meta-analysis but examined it in a separate analysis that does not include other study characteristics as covariates. The model used for aggregating the logits is the same as described earlier with only numeracy included as a covariate. The lack of other study characteristics as covariates means that the model does not account for large parts of heterogeneity. As a consequence, the amounts of residual heterogeneity are higher than in the original model ($\tau_F^2 = 1.5, \tau_P^2 = 1.0$) and the null hypothesis of no residual heterogeneity is rejected ($Q = 267.5, p < .0001$). The bottom panel of Table 2 shows the effects of numeracy. Ignoring differences due to other study characteristics, for participants with low numeracy scores it is estimated that 26 percent achieve the correct solution for the natural frequency format and 4 percent for the conditional probability format, resembling the baseline proportions in the full-sample meta-analysis. The performance of participants with high numeracy scores exceeds

those of participants with low scores by 25 percentage points in the natural frequency format and 11 percentage points in the conditional probability format. These effects appear large but ignore differences in study design and cannot be directly compared to effect estimates from the full-sample meta-analysis[11]. Additional studies on the effect of numeracy are required for other covariates be to included in such an analysis, and for reducing the current imprecision of the estimated effect of numeracy.

Nevertheless, our analysis shows that natural frequency formats continue to offer an advantage over conditional probability formats for low numerates and this difference is maintained for high numerates. These results support prior work showing a general format effect for natural frequencies (Hill & Brase, 2012; Johnson & Tubau, 2013) and do not suggest that the effect is stronger for high numerates (Chapman & Liu, 2009). Johnson and Tubau (2013) suggested that numeracy and information format can interact with the complexity of problems, for example in terms of information segmentation (e.g., in the gemstone problem described above) and verbal complexity, to impede performance for low numerates who may be more affected by such manipulations. Unfortunately, we were unable to examine other moderators used in these studies to draw any conclusions about interactions with verbal or computational complexity manipulations (Johnson & Tubau, 2013). There were also few studies that examined claims involving other measures of cognitive ability. In this view, Johnson and Tubau (2015) introduce a broader mathematical framework that takes into account individual differences in other areas, such as text comprehension and problem-solving abilities, to explore insights into solution strategies for Bayesian inference problems. We discuss this framework further below.

General Discussion

The results of the meta-analysis demonstrate that the natural frequency facilitation effect is fairly robust and is largely retained with respect to a range of individual, methodological, and problem representation moderators. Visual aids and short menu formats were the most influential moderators of the effect, enhancing performance for both natural frequency and conditional probability formats. These results suggest that not only is information structure an important

[11]There is, however, reason to assume that any effect estimated in this meta-analysis underestimates the effect of numeracy because all studies included here are based on median-splits in which group assignment to numeracy groups is based on the sample median score. When the distribution of test scores in the sample does not reflect the population distribution, this procedure may be biased because (some of) those below the sample median may fall above the population median. More accurate results can be obtained using "hard" cutoffs as are common in clinical research where assignment is based on pre-defined values. Such studies often require larger samples of participants.

component of the effect but that helping participants to *visualise* the information structure of the problem can improve performance. The computational complexity of the problems themselves also affects performance: adding additional cues or cue values does reduce performance rates across problems as anticipated, but adding an additional hypothesis can increase performance if the features of the problem allow one to ignore one or more of the hypotheses. The influence of methodological factors on performance rates, such as small improvements when individuals complete multiple problems or both formats, suggests that further insights could be gained from examining training or transfer effects across problems.

Limitations of the Meta-Analysis

Like all meta-analyses, the present meta-analysis is only as good as is the data available and most limitations of primary studies apply to their meta-analysis. For example, we were unable to analyse every moderator that has been tested across studies, as there were too few performance estimates for many potential moderators. For this reason, the meta-analysis does not do justice to the more subtle differences in study designs, although we attempted to control for the most important differences. Inevitably, there are more differences than the ones accounted for here and even among those, gaps in the available evidence lead to imprecise estimates.

Furthermore, some of the moderators included in the meta-analysis likely remain understudied. Given the variety of samples, problem representation manipulations, and methodologies utilised across studies, we had to collapse some of the coding categories as there was too little data to allow us to explore all potential coding categories in our analyses (e.g., adults versus children). In each case, we have attempted to justify collapsing coding categories given theoretical or methodological arguments found within the literature. We have also made sure to note caveats to any results involving the affected coding categories. Nonetheless, it remains possible that as a result subtle differences between these categories were overlooked.

Theory Building: Bridges or Fences?

The literature on the natural frequency facilitation effect has been enriched but also hindered by theoretical debates that foster the current dichotomy: ecological rationality framework versus nested-sets theory. In a recent theoretical review of Bayesian reasoning with natural frequencies, Brase and Hill (2015) criticised the persistence of the two "camps" into which many researchers place themselves, arguing that progress depends on researchers engaging in integration rather

than competition. Similarly, Johnson and Tubau (2015) have pushed for theory integration and proposed a framework for understanding Bayesian word problems that connects problem solving with mathematical cognition. In the following section, we discuss the implications of the results for the respective theories and make recommendations as to where further work can help to strengthen or clarify the premises of the theories or promote theory integration.

***Ecological* not evolutionary rationality framework.** The computational analysis of Bayesian inference problems originally put forward by Gigerenzer and Hoffrage (1995) is supported by the results of the meta-analysis in that short menu probability formats offered an advantage over standard conditional probability formats. Further, the improvement on short menu natural frequency formats suggests that simplifying even elementary arithmetic operations or specifying only the joint frequencies can facilitate the selection of relevant information. The results also suggest a role for visual displays in facilitating cognitive operations, a finding that connects to a vast literature on the development and use of visualisations for supporting thought (B. Tversky, 2001, 2011; Zahner & Corter, 2010). Although some proponents of the ecological rationality framework argue that specific types of visualisations may be most beneficial for Bayesian reasoning problems from an evolutionary perspective (see Brase, 2009b), others have not made specific predictions. We anticipate that proponents would argue that the more relevant question would be to ask which visual tools are most effective for different cognitive problems (see Zahner & Corter, 2010, for examples of different types of visual aids that are generated for different types of probability problems). Literature on the emergence of visual aids from a cultural or educational perspective may provide insights to guide research here.

Some of the criticism aimed at the ecological rationality framework as it has been applied to Bayesian reasoning is that the framework has not adequately addressed the question of how information about the occurrence of joint events is acquired or accumulated, nor has it offered a clear explanation for why some participants continue to have difficulty with natural frequency formats. This criticism could also be aimed at nested-sets theory. We turn to work on the *description-experience gap* (Hertwig, Barron, Weber, & Erev, 2004; Hertwig & Erev, 2009) to examine this criticism. Research on the description-experience gap emerged in consideration of contradictory findings related to how participants made decisions on the basis of probability distributions (Hertwig & Erev, 2009). When participants were given the opportunity to sample

event occurrences, from which probabilities and payoffs can be inferred, common errors found in studies that simply provide probabilities diminish or disappear. For example, base rates are more likely to be used when experienced rather than simply described (Koehler, 1996). Like others (Hoffrage, Krauss, et al., 2015; Schulze & Hertwig, 2016), we wonder whether natural frequencies offer an intermediate solution by improving on conditional probability formats such that the information is presented in a format that more closely resembles how the information is naturally acquired, but that stops short of providing people with the understanding of information structure that coincides with experience. For example, in line with related work on experience-based probability learning tasks, experience may facilitate the construction of a causal model (see, e.g., Sobel et al., 2004). We return to this aspect below when discussing future directions.

One of the most fundamental premises ofthe ecological rationality framework is the question of how cognitive processes map onto structures, or the *ecological* aspect of the framework (Gigerenzer & Hoffrage, 2007), yet this has received the least attention in research on Bayesian reasoning (see Gigerenzer & Hoffrage, 1995; Hafenbrädl & Hoffrage, 2015). Rather, much of the criticism of the theory resides in objections to an evolutionary argument regarding the potential for the human mind to have evolved a frequency-processing mechanism (Navarrete & Santamaria, 2011). This argument is not wholly supported by proponents of the theory, but nevertheless has become much more central to the debate than the ecological argument on which it is was originally based. At this point, we think it is pertinent to emphasise the response of Gigerenzer and Hoffrage (2007) to the many different interpretations of the ecological rationality framework put forward by Barbey and Sloman (2007):

> The evolutionary perspective ... provides a general framework for finding the right questions... An ecological framework postulates that thought does not simply emerge inside the mind. Every theory of reasoning needs to specify both cognitive strategies and the environmental structures under which these strategies work well (p. 266).

We wonder whether some of the debate between the ecological rationality framework and nested-sets theory would dissipate should the strict modularity view lose its emphasis.

Clarification of nested-sets theory. Given that nested-sets theory and the ecological rationality framework make similar arguments on the importance of the nested information structure to the facilitation effect of natural frequencies, the two theories can also draw on

similar results from the meta-analysis to support their premises. For example, the facilitative effect of short menu formats supports the premise of nested-sets theory that clarifying the nested-set structure of the problem can improve performance. Although the theory justifies the benefits of short menu formats on the basis of their ability to help participants visualise the subset structure, the mechanism by which this is revealed is not entirely clear. Mandel (2007) proposes that representations that reveal nested-sets and minimise computational complexity will enhance performance (holding transparency constant), which he refers to as the *complexity principle* of nested-sets theory. It is not clear how this principle differs from the computational argument made by Gigerenzer and Hoffrage (1995). Further, it is not clear how the theory accounts for the reduction in the natural frequency facilitation effect given short menu formats; the set structure is clarified in the same way in short menu versions of both formats.

Similarly, the prediction that visual aids can improve performance on conditional probability problems is supported in the present meta-analysis. Visual aids could enhance performance to a level similar to natural frequencies without visual aids. Given that visual aids also improved performance for natural frequency formats, we wonder whether the theory would argue that visual aids and natural frequency formats have an additive effect or whether the different methods for revealing subset structures build on one another. Mandel (2007) suggests that nested-sets relations can be clarified through different modalities (which he called the *multi-modal principle*), however, to our knowledge relations between these modalities has not been clarified. For instance, in relation to visual aids, the theory falls short of discussing whether the relative size or structure of the visual aid is important for clarifying the subsets (e.g, see Moro et al., 2011), or whether any visual design that illustrates subsets is sufficient.

One argument mounted in favour of nested-sets theory is that it has broader applications, such that transparent nested-set manipulations can facilitate *deductive* reasoning as well (Amitani, 2015; Barbey & Sloman, 2007). For example, Euler circles showing the subset structure of syllogisms can facilitate solutions (Sloman et al., 2003). Mandel (2007) alludes to the fact that there can be multiple ways in which the clarity of a representation can be improved, however, details of the range of strategies have not been elaborated. In fact, some proponents of the theory emphasise that *any* manipulation that draws one's attention to the nesting of events will facilitate reasoning (Lesage et al., 2013). Unfortunately, this claim allows the theory to accommodate a broad range of results in support of its premises without providing explanations for the mechanisms (whether these are the same or different across representations). For

example, in what ways can nested-set relations be made transparent? In which modalities? Is there a hierarchy of manipulations and how do different modalities operate in connection with one another? The theory needs to imply conditions for testing these mechanisms. On the other hand, the ecological rationality framework has also been criticised for not being able to explain how variations in the transparency of nested-sets, for example, textual manipulations, affect performance (Mandel, 2007).

At this point, we are inclined to agree with Mandel (2007) in his description of nested-sets theory as an assemblage of empirical findings collected in rebuttal to the frequentist mind perspective. Barbey and Sloman (2007) attempted to align the theory with dual-process models of reasoning to suggest that the deliberate rule-based system induces the use of rules about elementary set operations when set relations are transparent, although this dual process view is not wholly supported (Evans & Elqayam, 2007; Lagnado & Shanks, 2007; Mandel, 2007; Samuels, 2007). Sloman et al. (2003) allude to the fact that the ability to reason in relation to sets and subsets, including their relations and their relative sizes, is necessary for many problems (including those under primitive conditions, such as sharing resources) yet argues that claims of adaptiveness do not provide any further explanatory power. Yet, how did we come to operate so well on set relations? Nested-sets theory would benefit from further explication of its key propositions and principles, and from addressing the ecological nature of its predictions.

In summary, one of the major points of difference between the ecological rationality framework and nested-sets theory is the emphasis that evolutionary theory has in the foundations of the respective theories. Yet, it tends to be the proponents of nested-sets theory who continue to point to strong evolutionary claims within the ecological rationality framework, more so than its proponents. Rather, we think that a focus on the *ecological* aspect of the ecological rationality framework generates more interesting research questions. In any case, the two theoretical perspectives are broader than their application to Bayesian reasoning problems, which offer an environment for testing the predictions of the respective theories. However, if these predictions are ill-specified, this limits the insights that can be gained from this research.

Future Directions

The research reviewed in the present meta-analysis represents just a small part of research on Bayesian reasoning. While elementary Bayesian textbook problems have offered an experimental paradigm for testing theories, they represent a small world and this no doubt limits the breadth

of questions that can be addressed about Bayesian inference. Indeed, research within this paradigm led to the important insight that information representation matters to Bayesian reasoning and allowed researchers to explore the importance of information structure relative to other factors (e.g., numeracy, visual aids). However, we can still not offer a coherent explanation for why, in some cases, the majority of participants have difficulties with Bayesian reasoning problems as they have been studied here. Schulze and Hertwig (2016) suggest that differences in research methodology can help to explain some of these findings, for example, the finding that children are good intuitive statisticians but adults are not (e.g., employing experienced-based versus descriptive methods, respectively). We discuss ways in which research within this paradigm can be improved to generate further insights into the information representations that boost probabilistic inference.

Where and when do difficulties in Bayesian reasoning arise? In the reviewed studies, Bayesian reasoning has typically been defined as the ability to provide a correct probability estimate given the information provided. Focusing on this endpoint has limited our ability to understand how or determine why many of the interventions reviewed in our meta-analysis do or do not work (McNair, 2015). We discuss two opportunities to build on insights from research with Bayesian textbook problems to explore how representation can affect how people process the information and to explore differences across a range of performance criteria.

First, the use of a narrow criterion to evaluate performance on these descriptive tasks restricts arguments about Bayesian inference to a rather narrow mathematical definition and detracts from other potential questions (Domurat, Kowalczuk, Idzikowska, Borzymowska, & Nowak-Przygodzka, 2015; McNair, 2015; Wu et al., 2016). For example, given the different information formats, are people able to make the correct choice or inference given the information provided, irrespective of whether or not they can calculate a correct estimate? The results of Wu et al. (2016) and Domurat et al. (2015) suggest that making a correct choice does not necessarily imply that one calculates an exact estimate. Wu et al. (2016) employed an information selection task that required participants to select which of two tests had a better chance of answering a specific query (i.e., which of two genetic tests was most effective for identifying a species) presented either in natural frequency, conditional probability, or visual formats. The authors found no relation between probability judgement errors and choices, such that lower probability judgement errors were not necessarily associated with better choices.

Similarly, Domurat et al. (2015), employing a natural sampling approach for participants to learn probabilities, found that the majority of participants made choices that satisfied Bayes' theorem despite participants' verbal reports suggesting that non-Bayesian solution strategies were often applied. These results highlight how different performance criteria can lead to different conclusions about people's capabilities. It is an open question as to whether the information formats and the moderators examined in the present meta-analysis would have similar effects given different performance criteria.

Second, focusing on the difficulties that arise during the solution process can help identify the features of information representations that underlie the facilitation effect. Questions related to process have not played a central role in many studies despite the fact that Gigerenzer and Hoffrage (1995) included both performance and process measures in their original study. Their analysis of the visual analogs and solution strategies participants wrote down while solving Bayesian reasoning tasks helped them to identify cognitive shortcuts that could lead to solution rates similar to Bayes' theorem given specific features of the problem. Process measures represented an important part of their ecological analysis, to identify cognitive mechanisms and to identify when cognitive shortcuts could lead to correct solutions, for both conditional probability and natural frequency formats. However, only recently has there been renewed interest in the processes leading up to a correct solution rather than on endpoints alone (Brase & Hill, 2015; Johnson & Tubau, 2015; McNair, 2015).

For instance, Sirota, Vallée-Tourangeau, Vallée-Tourangeau, and Juanchich (2015), McNair (2015), and Johnson and Tubau (2015) have suggested that we draw on theories in the field of problem-solving and mathematical cognition for insights into how people approach and solve Bayesian reasoning problems. For example, Sirota, Vallée-Tourangeau, et al. (2015) suggest decomposing the question *"What facilitates Bayesian reasoning?"* into *"What facilitates the insight?"* or understanding of information structure, and *"What facilitates the computation?"*. Similarly, McNair (2015) suggests that focusing on process can help to identify cognitive abilities that are influential in early stages of the problem-solving process, and may indicate a lack of formal knowledge, and those that occur later and indicate a lack of ability to apply knowledge. Johnson and Tubau (2015) propose that text comprehension and problem solving are interrelated processes, and that understanding the relation between these processes is central to improve our understanding why and at which point of the process different intervention strategies, such as natural frequencies, work. In the following section we suggest how different

research methodologies offer opportunities to gain these insights.

Broadening methodological scope. A resounding criticism of work on Bayesian reasoning with elementary word problems is that these tasks are limited in what they can tell us about how people acquire, learn, or represent information about the occurrence of joint events (Brase & Hill, 2015; Girotto & Pighin, 2015; Mandel, 2014; Vallée-Tourangeau, Sirota, Juanchich, & Vallée-Tourangeau, 2015). For instance, in textbook tasks estimates are typically provided, and multiple estimates are not collected across time. Our understanding of probabilistic inference from descriptive tasks can be improved by incorporating different research methodologies, potentially leading to insights into theories and the assumptions on which they rest (as an example, consider research on the description-experience gap, Hertwig & Erev, 2009).

To test an assumption of the ecological rationality framework about how participants learn joint frequencies of events, Leach (2002) provided participants with either summary estimates of the relation between one of two symptoms and one of two diseases in a descriptive task, or with patient record cards that contained information about symptoms and diseases in an experiential natural sampling task. In both conditions, participants made highly accurate posterior probability estimates about the relation between symptoms and diseases. Similarly, Vallée-Tourangeau, Abadie, and Vallée-Tourangeau (2015) found that participants who received natural frequency or conditional probability text formats improved when these were accompanied by interactive cards displaying the joint occurrences of events, although performance was superior when the text represented natural frequencies. For both formats, the interactive cards allowed participants to see the subset structure or the relation between hypotheses and data. As far as we are aware, this is the only study that combines interactive sampling of joint events with a conditional probability descriptive information format. Experience-based methodologies could also be employed to explore updating or revision of posterior probabilities given new information.

Further, process tracing methods such as eye-tracking and verbal or written protocols can be employed to examine where attention is focused, to gauge the weight a reasoner is giving to different pieces of information, or to identify when in the process deviations occur (Johnson & Tubau, 2015; McNair, 2015). Another promising line of research is to assess the prior distributions people have for different types of real events (Mandel, 2014). In this connection, Obrecht, Anderson, Schulkin, and Chapman (2012) asked obstetricians and gynaecologists to estimate the probability of Down syndrome given a positive test result and subsequently, to report esti-

mates of the base rate (babies with Down syndrome), hit rate (positive test result given Down syndrome), and false-alarm rate (positive test result given no Down syndrome) in either natural frequencies or conditional probabilities. Cognisant of differences in the physician's personal experiences, Obrecht et al. (2012) examined accuracy in terms of whether the participants' base rate, hit rate, and false-alarm rate estimates were consistent with their posterior probability estimate. Posterior probability estimates were more consistent when information was requested in natural frequencies as opposed to conditional probabilities, a finding that generates questions about how people store information about the occurrence of joint events. There are many opportunities for research on Bayesian reasoning to broaden its methodological scope and explore a variety of research questions on information structure and reasoning through descriptive tasks.

From Textbook Problems to Real World Applications

At this point, it is important to remember that the study of Bayesian reasoning using word problems emerged in consideration of real world contexts where information is communicated in similar formats (e.g., Eddy, 1982). As Navarrete, Correia, Sirota, Juanchich, and Huepe (2015) emphasise, we should not lose sight of the ultimate goal to foster understanding in contexts where probabilistic inference from description problems is required. A few studies have sought to examine the effect of information representation formats in groups who are required to make these types of probabilistic inferences in a given domain, for example: medical professionals (Ben-Shlomo, Collin, Quekett, Sterne, & Whiting, 2015; Bramwell et al., 2006; Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007; Hoffrage & Gigerenzer, 1998), managers (Hoffrage, Hafenbrädl, & Bouquet, 2015), and advanced law students and professional jurists (Lindsey, Hertwig, & Gigerenzer, 2003). Attempts to improve probabilistic inferences through training participants to translate probabilities into natural frequency representations (as opposed to rule-based training using Bayes' rule) have generally been shown to be effective over the longer term (Kurzenhäuser & Hoffrage, 2002; Sedlmeier & Gigerenzer, 2001; Sirota, Kostovičová, & Vallée-Tourangeau, 2015a). However, we are unaware of more formalised efforts to implement training in curricula. Further, a better understanding of the errors participants make with each format and the implications of these errors on inference is needed. For example, future work could focus on the different cognitive shortcuts participants make and identify the conditions under which these shortcuts can approximate correct solutions (e.g., see Gigerenzer & Hoffrage, 1995; Hafenbrädl & Hoffrage, 2015).

Conclusions

The facilitative effect of natural frequencies is robust and our meta-analysis identified conditions under which performance on conditional probability problems can be improved further. Thus, although there remains room for improvement on natural frequency formats, the results of the meta-analysis suggest that natural frequencies are favourable to conditional probability formats. Even though short menu formats and visual aids can improve performance in both natural frequency and conditional probability formats, these moderators are still better used with natural frequency formats, with visual aids offering the strongest advantage to performance. We had hoped to examine the relative benefits of different visual designs but were limited by the number of studies that have explored different design features and thus, further work is needed to establish which visual aids are most effective and why (although see Böcherer-Linder & Eichler, 2016 Wu et al., 2016, and Khan et al., 2015, for exceptions).

There is also preliminary evidence to suggest that higher numeracy and performance-based incentives can improve performance on Bayesian inference tasks. However, there is a lack of comparative studies examining these moderators, as well as non-university samples (e.g., children or experts), and textual manipulations aimed at improving problem comprehension (as opposed to texts aimed to emphasise set relations). The meta-analysis also identified how variations in performance could be explained by differences in study designs, such as when participants complete both problem formats. We suggest that current research methodologies employed in the study of elementary textbook tasks can be extended to incorporate more experience-based and process-tracing approaches. However, what we have learned from the many studies included in our review is that not only can natural frequencies improve Bayesian inference, but that there is still ample room for improvement. We hope that future work will focus not only on different performance criteria but on the processes leading up to the correct solution as well, with the aim to understand why many participants continue to have difficulties solving Bayesian inference tasks. Ultimately, future work will move beyond the current theoretical dichotomy of the ecological rationality framework and nested-sets theory to focus on integration, not only between but also beyond these two perspectives.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Amitani, Y. (2015). The natural frequency hypothesis and evolutionary arguments. *Mind & Society*, *14*(1), 1-19. doi: 10.1007/s11299-014-0155-7

Ayal, S., & Beyth-Marom, R. (2014). The effects of mental steps and compatibility on Bayesian reasoning. *Judgment and Decision Making*, *9*, 226-242.

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

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*, 211-233. doi: 10.1016/0001-6918(80)90046-3

Bar-Hillel, M. (1984). Representativeness and fallacies of probability judgment. *Acta Psychologica*, *55*, 91-107. doi: 10.1016/0001-6918(84)90062-3

Barton, A., Mousavi, S., & Stevens, J. R. (2007). A statistical taxonomy and another "chance" for natural frequencies. *Behavioral and Brain Sciences*, *30*, 255-256. doi: 10.1017/S0140525X07001665

Ben-Shlomo, Y., Collin, S. M., Quekett, J., Sterne, J. A., & Whiting, P. (2015). Presentation of diagnostic information to doctors may change their interpretation and clinical management: A web-based randomised controlled trial. *PLoS One*, *10*(7), e0128637. doi: 10.1371/journal.pone.0128637

Biernaskie, J. M., Walker, S. C., & Gegear, R. J. (2009). Bumblebees learn to forage like Bayesians. *The American Naturalist*, *174*(3), 413-423. doi: 10.1086/603629

*Binder, K., Krauss, S., & Bruckmaier, G. (2015). Effects of visualizing statistical information: An empirical study on tree diagrams and 2 × 2 tables. *Frontiers in Psychology*, *6*, 1-9. doi: 10.3389/fpsyg.2015.01186

Böcherer-Linder, K., & Eichler, A. (2016). The impact of visualizing nested sets: An empirical study on tree diagrams and unit squares. *Frontiers in Psychology*, *7*, 2026. doi: 10.3389/fpsyg.2016.02026

*Bramwell, R., West, H., & Salmon, P. (2006). Health professionals' and service users' interpretation of screening test results: Experimental study. *BMJ*, *333*, 284–286A. doi: 10.1136/bmj.38884.663102.AE

Brase, G. L. (2007). Omissions, conflations, and false dichotomies: Conceptual and empirical problems with the Barbey & Sloman account. *Behavioral and Brain Sciences*, *30*, 258-259. doi: 10.1017/S0140525X07001690

*Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information fa-

cilitates Bayesian reasoning. *Psychonomic Bulletin & Review*, *15*, 284-289. doi: 10.3758/PBR15.2.284

*Brase, G. L. (2009a). How different types of participant payments alter task performance. *Judgment and Decision Making*, *4*(5), 419-428.

Brase, G. L. (2009b). Pictorial representations in statistical reasoning. *Applied Cognitive Psychology*, *23*, 369-381. doi: 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. doi: 10.1080/20445911.2013.861840

Brase, G. L., Fiddick, L., & Harries, C. (2006). Participant recruitment methods and statistical reasoning performance. *Quarterly Journal of Experimental Psychology*, *59*, 965-976. doi: 10.1080/0272498054300132

Brase, G. L., & Hill, W. T. (2015). Good fences make for good neighbors but bad science: A review of what improves Bayesian reasoning and why. *Frontiers in Psychology*, *6*, 1-9. doi: 10.3380/fpsyg.2015.00340

Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin*, *140*(4), 980-1008. doi: 10.1037/a0035661

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

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 811-23. doi: 10.1002/wcs.79

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287-91. doi: 10.1016/j.tics.2006.05.007

Cole, W. G. (1988). Three graphic representations to aid Bayesian inference. *Methods of Information in Medicine*, *27*, 125-132.

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. doi: 10.1016/0010-0277(95)00664-8

Cosmides, L., & Tooby, J. (2008). Can a general deontic logic capture the facts of human moral reasoning? How the mind interprets social exchange rules and detects cheaters. In

W. Sinnott-Armstrong (Ed.), *Moral psychology* (p. 53-119). Cambridge, MA: MIT Press.

Domurat, A., Kowalczuk, O., Idzikowska, K., Borzymowska, Z., & Nowak-Przygodzka, M. (2015). Bayesian probability estimates are not necessary to make choices satisfying Bayes' rule in elementary situations. *Frontiers in Psychology*, *6*, 1-14. doi: 10.3389/fpsyg.2015.01194

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. doi: 10.1111/j.0006-341X.2000.00455.x

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

Evans, J. S. B. T., & Elqayam, S. (2007). Dual-processing explains base-rate neglect, but which dual-process theory and how? *Behavioral and Brain Sciences*, *30*, 261-262. doi: 10.1017/S0140525X07001720

*Evans, J. S. B. T., Handley, S. J., Perham, N., Over, D. E., & Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. *Cognition*, *77*, 197-213. doi: 10.1016/S0010-0277(00)00098-6

*Ferguson, E., & Starmer, C. (2013). Incentives, expertise, and medical decisions: Testing the robustness of natural frequency framing. *Health Psychology*, *32*, 967-977. doi: 10.1037/a0033720

*Fiedler, K., Brinkmann, B., Betsch, T., & Wild, B. (2000). A sampling approach to biases in conditional probability judgments: Beyond base rate neglect and statistical format. *Journal of Experimental Psychology: General*, *129*, 399-418. doi: 10.1037/0096-3445.129.3.399

Fiedler, K., & von Sydow, M. (2015). Heuristics and biases: Beyond Tversky and Kahneman's (1974) judgment under uncertainty. In M. W. Eysenck & D. Groome (Eds.), *Cognitive psychology: Revisiting the classical studies* (p. 146-161). Los Angeles, CA: Sage.

Fontanari, L., Gonzalez, M., Vallortigara, G., & Girotto, V. (2014). Probabilistic cognition in two indigenous mayan groups. *Proceedings of the National Academy of Sciences*, *111*, 17075-17080. doi: 10.1073/pnas.1410583111

*Friederichs, H., Ligges, S., & Weissenstein, A. (2014). Using tree diagrams without numerical values in addition to relative numbers improves students' numeracy skills: A

randomized study in medical education. *Medical Decision Making*, *34*, 253-257. doi: 10.1177/0272989X13504499

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology*, *28*(2), 210-216. doi: 10.1037/a0014474

*Galesic, M., Gigerenzer, G., & Straubinger, N. (2009). Natural frequencies help older adults and people with low numeracy to evaluate medical screening tests. *Medical Decision Making*, *29*, 368-371. doi: 10.1177/0272989X08329463

Garcia-Retamero, R., & Cokely, E. T. (2013). Communicating health risks with visual aids. *Current Directions in Psychological Science*, *22*(5), 392-399. doi: 10.1177/0963721413491570

Garcia-Retamero, R., & Hoffrage, U. (2013). Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Social Science & Medicine*, *83*, 27-33. doi: 10.1016/j.socscimed.2013.01.034

Gigerenzer, G. (1996a). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, *103*, 592-596. doi: 10.1037/0033-295X.103.3.592

Gigerenzer, G. (1996b). Why do frequency formats improve Bayesian reasoning? Cognitive algorithms work on information, which needs representation. *Behavioral and Brain Sciences*, *19*, 23-24. doi: 10.1017/S0140525X00041248

Gigerenzer, G. (1998). Ecological intelligence: An adaptation for frequencies. In D. D. Cummins & C. Allen (Eds.), *The evolution of mind* (p. 9-29). New York, NY: Oxford University Press.

Gigerenzer, G. (2015). On the supposed evidence for libertarian paternalism. *Review of Philosophy and Psychology*, *6*(3), 361-383. doi: 10.1007/s13164-015-0248-1

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*(2), 53-96. doi: 10.1111/j.1539-6053.2008.00033.x

Gigerenzer, G., & Goldstein, D. G. (2011). The recognition heuristic: A decade of research. *Judgment and Decision Making*, *6*(1), 100-121.

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

Gigerenzer, G., & Hoffrage, U. (2007). The role of representation in Bayesian reasoning:

Correcting common misconceptions. *Behavioral and Brain Sciences*, *30*, 264-267. doi: 10.1017/S0140525X07001756

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability theory changed science and everyday life.* Cambridge, UK: Cambridge University Press.

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

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

Girotto, V., & Gonzalez, M. (2001). Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition*, *78*, 247-276. doi: 10.1016/S0010-0277(00)00133-5

Girotto, V., & Gonzalez, M. (2002). Chances and frequencies in probabilistic reasoning: Rejoinder to Hoffrage, Gigerenzer, Krauss, and Martignon. *Cognition*, *84*, 353-359. doi: 10.1016/S0010-0277(02)00051-3

Girotto, V., & Pighin, S. (2015). Basic understanding of posterior probability. *Frontiers in Psychology*, *6*, 680. doi: 10.3389/fpsyg.2015.00680

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, *37*, 620-629. doi: 10.1037//0012-1649.37.5.620

Grüne-Yanoff, T., & Hertwig, R. (2016). Nudge versus boost: How coherent are policy and theory? *Minds and Machines*, *26*(1), 149-183. doi: 10.1007/s11023-015-9367-9

Hacking, I. (2006). *The emergence of probability* (2nd ed.). 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*, 1-15. doi: 10.3389/fpsyg.2015.00939

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. doi: 10.1002/jrsm.5

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis.

Psychological Methods, *3*(4), 486. doi: 10.1037/1082-989X.3.4.486

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61-83. doi: 10.1017/S0140525X0999152X

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(8), 534-9. doi: 10.1111/j.0956-7976.2004.00715.x

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517-523. doi: 10.1016/j.tics.2009.09.004

Hertwig, R., Hoffrage, U., & the ABC Research Group. (2013). *Simple heuristics in a social world.* New York, NY: Oxford University Press.

Higgins, J., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. doi: 10.1002/sim.1186

Higgins, J., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*(7414), 557–560. doi: 10.1136/bmj.327.7414.557

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

*Hill, W. T., & Brase, G. L. (2015). *Natural frequencies improve diagnostic test result comprehension when using one, two, and three cues.* Unpublished manuscript.

*Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, *73*, 538-540. doi: 10.1097/00001888-199805000-00024

*Hoffrage, U., & Gigerenzer, G. (2004). How to improve the diagnostic inferences of medical experts. In E. Kurz-Milcke & G. Gigerenzer (Eds.), *Experts in science and society.* (p. 249-268). New York, NY: Kluwer Academic/Plenum Publishers. doi: 10.1007/b105826

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition*, *84*, 343-352. doi: 10.1016/S0010-0277(02)00050-1

*Hoffrage, U., Hafenbrädl, S., & Bouquet, C. (2015). Natural frequencies facilitate diagnostic inferences of managers. *Frontiers in Psychology*, *6*. doi: 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*, 1-14. doi: 10.3389/fpsyg.2015.01473

*Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science*, *290*, 2261-2262. doi: 10.1126/science.290.5500.2261

Jackson, D., White, I. R., & Riley, R. D. (2012). Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Statistics in Medicine*, *31*(29), 3805–3820. doi: 10.1002/sim.5453

*Johnson, E. D., & Tubau, E. (2013). Words, numbers, & numeracy: Diminishing individual differences in Bayesian reasoning. *Learning and Individual Differences*, *28*, 34-40. doi: 10.1016/j.lindif.2013.09.004

Johnson, E. D., & Tubau, E. (2015). Comprehension and computation in Bayesian problem solving. *Frontiers in Psychology*, *6*, 1-19. doi: 10.3389/fpsyg.2015.00938

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J. P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review*, *106*, 62-88. doi: 10.1037/0033-295X.106.1.62

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430-454. doi: 10.1016/0010-0285(72)90016-3

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237-251. doi: 10.1037/h0034747

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition*, *11*, 123-141. doi: 10.1016/0010-0277(82)90022-1

Khan, A., Breslav, S., Glueck, M., & Hornbaek, K. (2015). Benefits of visualization in the mammography problem. *International Journal of Human-Computer Studies*, *83*, 94-113. doi: 10.1016/j.ijhcs.2015.07.001

Kleiter, G. D. (1994). Natural sampling: Rationality without base rates. In G. H. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (p. 375-388). New York, NY: Springer. doi: 10.1007/978-1-4612-4308-3_27

Kochetova-Kozloski, N., Messier, W. F. J., & Eilifsen, A. (2011). Improving auditors' fraud judgments using a frequency response mode. *Contemporary Accounting Research*, *28*, 837-858. doi: 10.1111/j.1911-3846.2011.01067.x

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, *19*, 1-17. doi: 10.1017/S0140525X00041157

*Konheim-Kalkstein, Y. L. (2008). *Facilitation of Bayesian decision making.* (Doctoral disser-

tation). University of Minnesota, Twin Cities, MN.

*Krauss, S., Martignon, L., & Hoffrage, U. (1999). Simplifying Bayesian inference: The general case. In L. Magnani, N. J. Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery* (p. 165-179). New York, NY: Kluwer Academic/Plenum Publishers.

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, *136*, 430-450. doi: 10.1037/0096-3445.136.3.430

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

Lagnado, D. A., & Shanks, D. R. (2007). Dual concerns with the dualist approach. *Behavioral and Brain Sciences*, *30*, 271-272. doi: 10.1017/S0140525X0700180X

Lau, J., Ioannidis, J. P., Terrin, N., Schmid, C. H., & Olkin, I. (2006). Evidence based medicine: The case of the misleading funnel plot. *BMJ*, *333*(7568), 597-600. doi: 10.1136/bmj.333.7568.597

Leach, J. R. (2002). *Information selection in a simulated medical diagnosis task: The effects of external representations and compeltely natural sampling.* (Doctoral dissertation). Bowling Green State University, Bowling Green, OH.

*Lesage, E., Navarrete, G., & De Neys, W. (2013). Evolutionary modules and Bayesian facilitation: The role of general cognitive resources. *Thinking & Reasoning*, *19*, 27-53. doi: 10.1080/13546783.2012.713177

*Lindsey, S., Hertwig, R., & Gigerenzer, G. (2003). Communicating statistical DNA evidence. *Jurimetrics*, *43*(2), 147-163.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, *21*(1), 37-44. doi: 10.1177/0272989x0102100105

*Macchi, L. (2000). Partitive formulation of information in probabilistic problems: Beyond heuristics and frequency format explanations. *Organizational Behavior and Human Decision Processes*, *82*, 217-236. doi: 10.1006/obhd.2000.2895

Mandel, D. R. (2007). Nested sets theory, full stop: Explaining performance on Bayesian inference tasks without dual-systems assumptions. *Behavioral and Brain Sciences*, *30*, 275-276. doi: 10.1017/S0140525X07001835

Mandel, D. R. (2014). The psychology of Bayesian reasoning. *Frontiers In Psychology*, *5*, 1144. doi: 10.3389/fpsyg.2014.01144

Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review*, *118*(3), 393-437. doi: 10.1037/a0024143

McNair, S. (2015). Beyond the status-quo: Research on Bayesian reasoning must develop in both theory and method. *Frontiers in Psychology*, *6*, 1-3. doi: 10.3389/fpsyg.2015.00097

McNair, S., & Feeney, A. (2014). When does information about causal structure improve statistical reasoning? *Quarterly Journal of Experimental Psychology*, *67*, 625-645. doi: 10.1080/17470218.2013.821709

McNair, S., & Feeney, A. (2015). Whose statistical reasoning is facilitated by a causal structure intervention? *Psychonomic Bulletin & Review*, *22*(1), 258-264. doi: 10.3758/s13423-014-0645-y

*Mellers, B. A., & McGraw, A. P. (1999). How to improve Bayesian reasoning: Comment on Gigerenzer and Hoffrage (1995). *Psychological Review*, *106*, 417-424. doi: 10.1037/0033-295X.106.2.417

Micallef, L., Dragicevic, P., & Fekete, J.-D. (2012). Assessing the effect of visualizations on Bayesian reasoning through crowdsourcing. *IEEE Transactions On Visualization and Computer Graphics*, *18*, 2536-2545. doi: 10.1109/TVCG.2012.199

*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. doi: 10.2466/PMS.108.2.422-430

*Moro, R., Bodanza, G. A., & Freidin, E. (2011). Sets or frequencies? How to help people solve conditional probability problems. *Journal of Cognitive Psychology*, *23*, 843-857. doi: 10.1080/20445911.2011.579072

Navarrete, G., Correia, R., Sirota, M., Juanchich, M., & Huepe, D. (2015). Doctor, what does my positive test mean? From Bayesian textbook tasks to personalized risk communication. *Frontiers in Psychology*, *6*, 1-6. doi: 10.3389/fpsyg.2015.01327

Navarrete, G., & Santamaria, C. (2011). Ecological rationality and evolution: The mind really works that way? *Frontiers In Psychology*, *2*, 251. doi: 10.3389/fpsyg.2011.00251

Neace, W. P., Michaud, S., Bolling, L., Deer, K., & Zecevic, L. (2008). Frequency formats, probability formats, or problem structure? A test of the nested-sets hypothesis in an extensional reasoning task. *Judgment and Decision Making*, *3*, 140-152.

Obrecht, N. A., Anderson, B., Schulkin, J., & Chapman, G. B. (2012). Retrospective frequency formats promote consistent experience-based Bayesian judgments. *Applied Cognitive Psychology*, *26*, 436-440. doi: 10.1002/acp.2816

*O'Brien, D., Roazzi, A., & da Graca B. B. Dias, M. (2004). Reasoning about conditional probabilities: The evidence for the frequency hypothesis has relied on flawed comparisons. *Estudos de Psicologia*, *9*, 35-43. doi: 10.1590/S1413-294X2004000100005

Ottley, A., Peck, E. M., Harrison, L. T., Afergan, D., Ziemkiewicz, C., Taylor, H. A., ... Chang, R. (2016). Improving Bayesian reasoning: The effects of phrasing, visualization, and spatial ability. *IEEE Transactions On Visualization and Computer Graphics*, *22*, 529-538. doi: ieeecomputersociety.org/10.1109/TVCG.2015.2467758

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *JAMA*, *295*(6), 676–680. doi: 10.1001/jama.295.6.676

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*(1), 29-46. doi: 10.1037/h0024722

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*, 346-354. doi: 10.1037/h0023653

R Core Team. (2014). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rakoczy, H., Cluever, A., Saucke, L., Stoffregen, N., Graebener, A., Migura, J., & Call, J. (2014). Apes are intuitive statisticians. *Cognition*, *131*, 60-68. doi: 10.1016/j.cognition.2013.12.011

Real, L. (1991). Animal choice behavior and the evolution of cognitive architecture. *Science*, *253*(5023), 980-986. doi: 10.1126/science.1887231

Real, L., & Caraco, T. (1986). Risk and foraging in stochastic environments. *Annual Review of Ecology and Systematics*, *17*, 371-390. doi: 10.1146/annurev.es.17.110186.002103

Ruscio, J. (2003). Comparing Bayes's theorem to frequency-based approaches to teaching Bayesian reasoning. *Teaching of Psychology*, *30*, 325-328.

Samuels, R. (2007). Varieties of dual-process theory for probabilistic reasoning. *Behavioral and Brain Sciences*, *30*, 280-281. doi: 10.1017/S0140525X07001884

Savage, L. J. (1954). *The foundations of statistics.* New York, NY: Wiley.

Schulze, C., & Hertwig, R. (2016). *Statistical intuitions: Smart babies, stupid adults?*

Manuscript submitted for publication.

Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine*, *127*(11), 966-72. doi: 10.7326/0003-4819-127-11-199712010-00003

*Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130*, 380-400. doi: 10.1037//0096-3445.130.3.380

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology*, *62*(4), 273-296. doi: 10.1016/j.cogpsych.2011.03.001

*Siegrist, M., & Keller, C. (2011). Natural frequencies and Bayesian reasoning: The impact of formal education and problem context. *Journal of Risk Research*, *14*, 1039-1055. doi: 10.1080/13669877.2011.571786

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

*Sirota, M., Juanchich, M., & Hagmayer, Y. (2014). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychonomic Bulletin & Review*, *21*, 198-204. doi: 10.3758/s13423-013-0464-6

Sirota, M., Kostovičová, L., & Juanchich, M. (2014). The effect of iconicity of visual displays on statistical reasoning: Evidence in favor of the null hypothesis. *Psychonomic Bulletin & Review*, *21*, 961-8. doi: 10.3758/s13423-013-0555-4

Sirota, M., Kostovičová, L., & Vallée-Tourangeau, F. (2015a). How to train your Bayesian: A problem-representation transfer rather than a format-representation shift explains training effects. *Quarterly Journal of Experimental Psychology*, *68*, 1-9. doi: 10.1080/17470218.2014.972420

Sirota, M., Kostovičová, L., & Vallée-Tourangeau, F. (2015b). Now you Bayes, now you don't: Effects of set-problem and frequency-format mental representations on statistical reasoning. *Psychonomic Bulletin & Review*, *22*(5), 1465-1473. doi: 10.3758/s13423-015-0810-y

Sirota, M., Vallée-Tourangeau, G., Vallée-Tourangeau, F., & Juanchich, M. (2015). On Bayesian problem-solving: Helping Bayesians solve simple Bayesian word problems. *Frontiers in Psychology*, *6*, 1-4. doi: 10.3389/fpsyg.2015.01141

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

Sobel, D. M., & Munro, S. E. (2009). Domain generality and specificity in children's causal inference about ambiguous data. *Developmental Psychology*, *45*, 511-524. doi: 10.1037/a0014944

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, *28*, 303-333. doi: 10.1016/j.cogsci.2003.11.001

Stanovich, K. E., & West, R. F. (2000). Advancing the rationality debate. *Behavioral and Brain Sciences*, *23*(05), 701–717. doi: 10.1017/S0140525X00623439

Stedman, M. R., Curtin, F., Elbourne, D. R., Kesselheim, A. S., & Brookhart, M. A. (2011). Meta-analyses involving cross-over trials: Methodological issues. *International Journal of Epidemiology*, *40*(6), 1732–1734. doi: 10.1093/ije/dyp345

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press.

Todd, P. M., & Brighton, H. (2015). Building the theory of ecological rationality. *Minds and Machines*, 1-22. doi: 10.1007/s11023-015-9371-0

Todd, P. M., & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science*, *16*(3), 167-171. doi: 10.1111/j.1467-8721.2007.00497.x

Todd, P. M., Gigerenzer, G., & the ABC Research Group. (2012). *Ecological rationality: Intelligence in the world*. New York, NY: Oxford University Press Inc.

Todd, P. M., Hertwig, R., & Hoffrage, U. (2005). Evolutionary cognitive psychology. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (p. 776-802). Hoboken, NJ: Wiley.

Trafimow, D. (2007). Why the empirical literature fails to support or disconfirm modular or dual-process models. *Behavioral and Brain Sciences*, *30*, 283-284. doi: 10.1017/S0140525X07001926

*Tsai, J., Miller, S., & Kirlik, A. (2011). Interactive visualizations to improve Bayesian reasoning. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 385–389. doi: 10.1177/1071181311551079

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293-315. doi: 10.1037/0033-295X.90.4.293

Tversky, B. (2001). Spatial schemas in depictions. In M. Gattis (Ed.), *Spatial schemas and abstract thought* (p. 79-112). Massachusetts, MA: MIT Press.

Tversky, B. (2011). Visualizing thought. *Topics in Cognitive Science*, *3*(3), 499-535. doi: 10.1111/j.1756-8765.2010.01113.x

*Vallée-Tourangeau, G., Abadie, M., & Vallée-Tourangeau, F. (2015). Interactivity fosters Bayesian reasoning without instruction. *Journal of Experimental Psychology: General*, *144*(3), 581-603. doi: 10.1037/a0039161

Vallée-Tourangeau, G., Abadie, M., & Vallée-Tourangeau, F. (2016). *Interactivity fosters Bayesian reasoning without instruction* [Data file and code book]. Retrieved from `osf.io/2ur8f`

Vallée-Tourangeau, G., Sirota, M., Juanchich, M., & Vallée-Tourangeau, F. (2015). Beyond getting the numbers right: What does it mean to be a "successful" Bayesian reasoner? *Frontiers in Psychology*, *6*, 712. doi: 10.3389/fpsyg.2015.00712

van Assen, M. A., van Aert, R., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, *20*(3), 293-309. doi: 10.1037/met0000025

van Houwelingen, H. C., Zwinderman, K. H., & Stijnen, T. (1993). A bivariate approach to meta-analysis. *Statistics in Medicine*, *12*, 2273–2284. doi: 10.1002/sim.4780122405

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48. doi: 10.18637/jss.v036.i03

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125. doi: 10.1002/jrsm.11

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of Human Genetics*, *19*(4), 251–253. doi: 10.1111/j.1469-1809.1955.tb01348.x

Wu, C. M., Meder, B., Nelson, J. D., & Filimon, F. (2016). *Asking better questions: How presentation formats influence information search.* Manuscript submitted for publication.

*Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: Frequency

or nested sets? *Experimental Psychology*, *50*, 97-106. doi: 10.1026//1618-3169.50.2.97

Zahner, D., & Corter, J. E. (2010). The process of probability problem solving: Use of external visual representations. *Mathematical Thinking and Learning*, *12*(2), 177-204. doi: 10.1080/10986061003654240

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

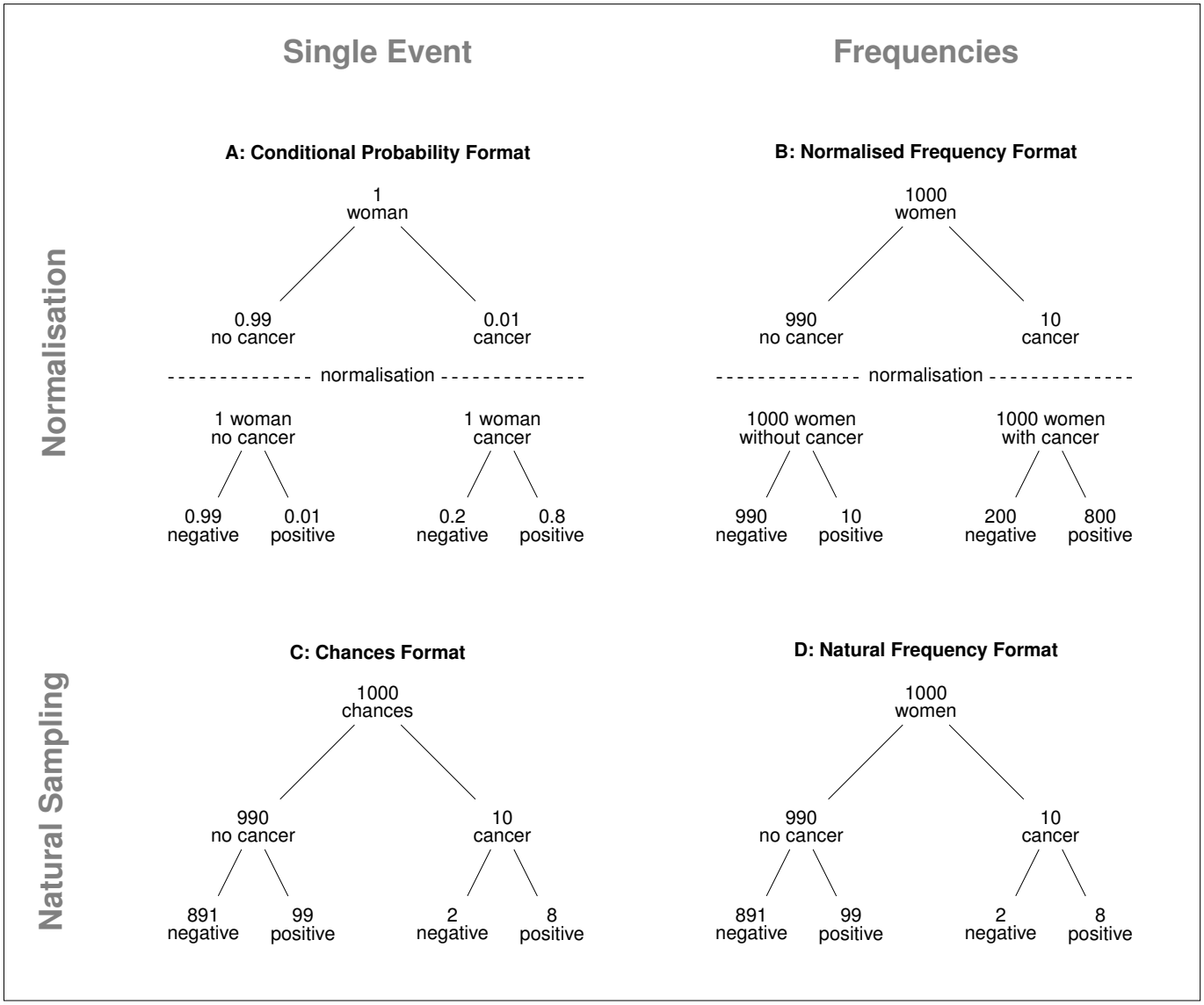

Figure 1: Taxonomy of representation formats (adapted from Gigerenzer & Hoffrage, 2007).

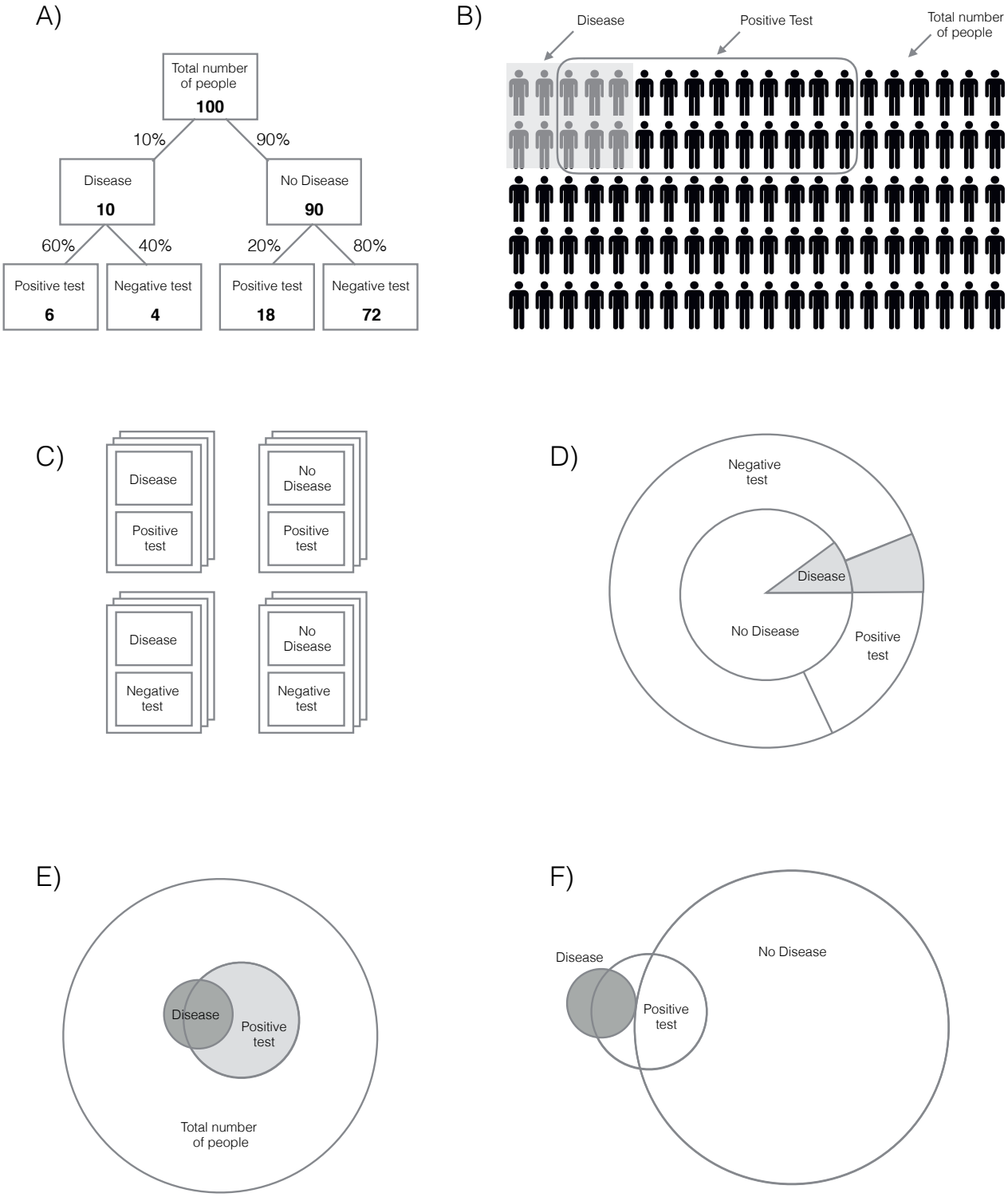

Figure 2: Examples of visual aids for conditional probability and natural frequency problems. A) Natural frequency or conditional probability tree (see, e.g., Binder et al., 2015; Sedlmeier & Gigerenzer, 2001). Numbers in tree represent natural frequencies and numbers beside branches represent probability versions. B) Icon array/frequency grid (e.g., Brase, 2009a, 2014). C) Interactive cards (e.g., Vallée-Tourangeau, Abadie, & Vallée-Tourangeau, 2015). D) Roulette wheel (e.g., Brase, 2014; Yamagishi, 2003). E) Euler diagram and F) Area proportional Euler diagram (e.g., Micallef et al., 2012; Sloman et al., 2003).

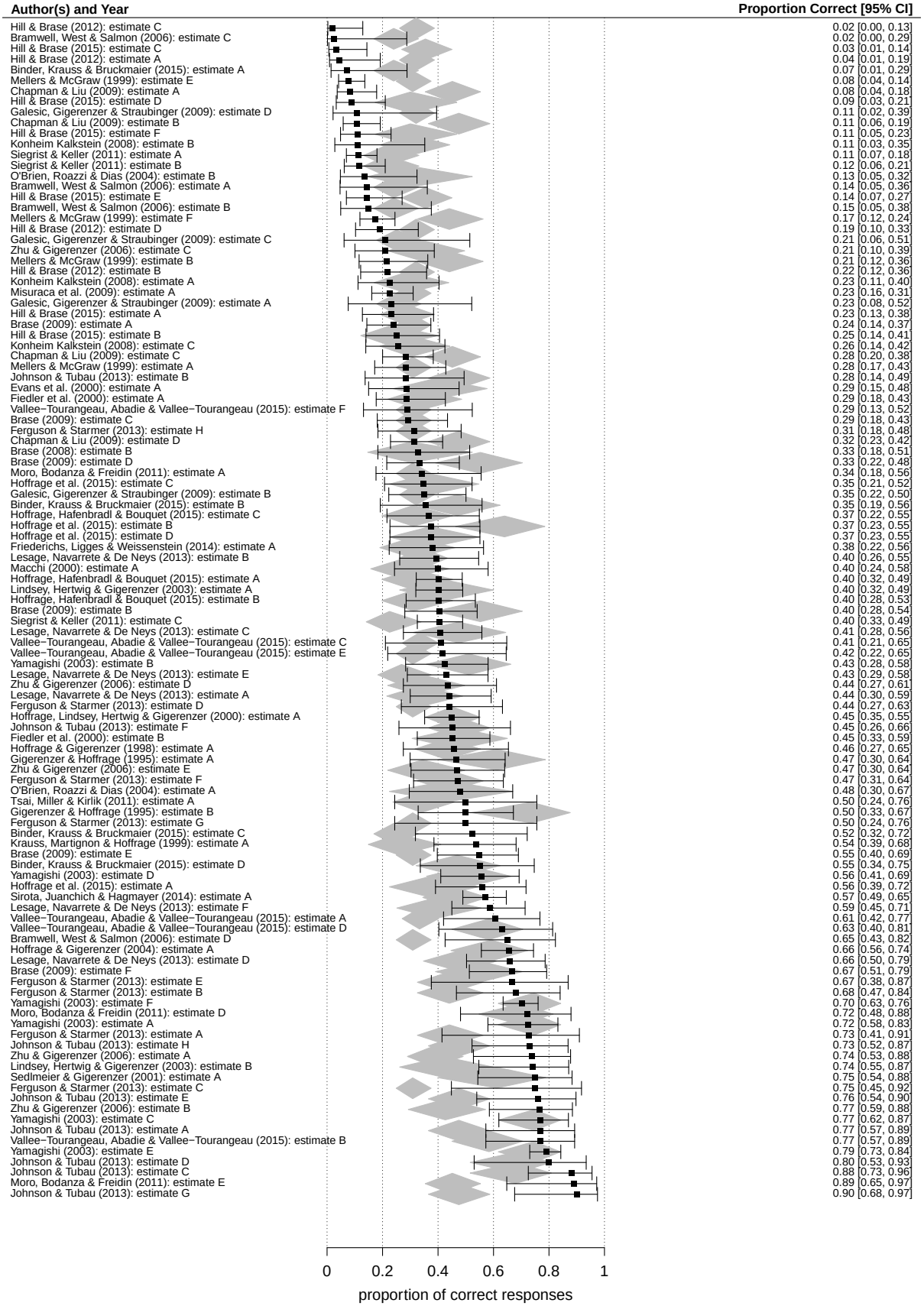

Figure 3: Forest plot of proportions correct in natural frequency format: The black squares represent the observed values and the whiskers their corresponding confidence intervals. The polygons represent the estimated average proportions (and their confidence intervals) for studies with the same characteristics. Estimates A, B, C,... denote different comparisons from the same paper, referenced in Table 3.

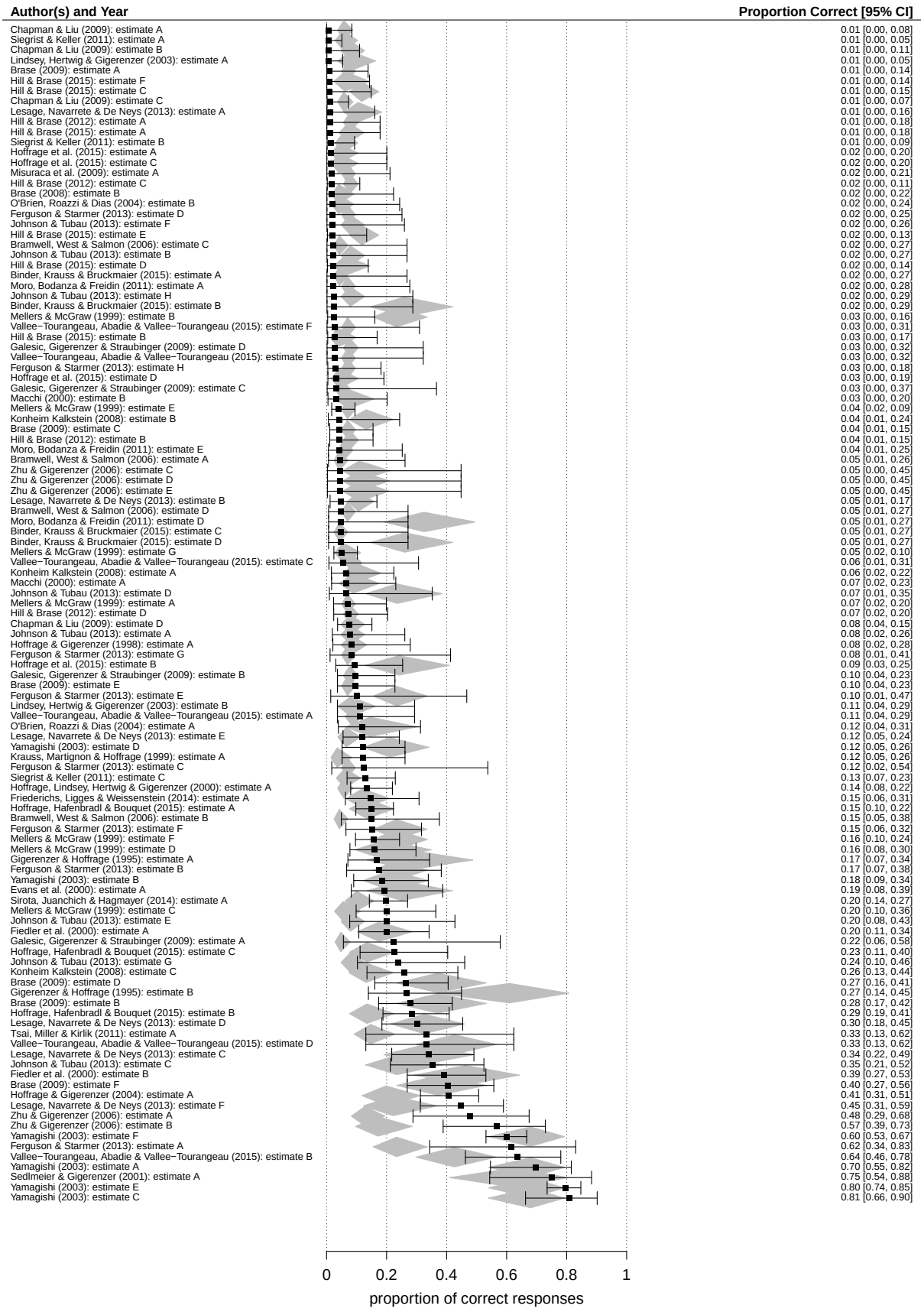

Figure 4: Forest plot of proportions correct in conditional probability format: The black squares represent the observed values and the whiskers their corresponding confidence intervals. The polygons represent the estimated average proportions (and their confidence intervals) for studies with the same characteristics. Estimates A, B, C,... denote different comparisons from the same paper, referenced in Table 3.

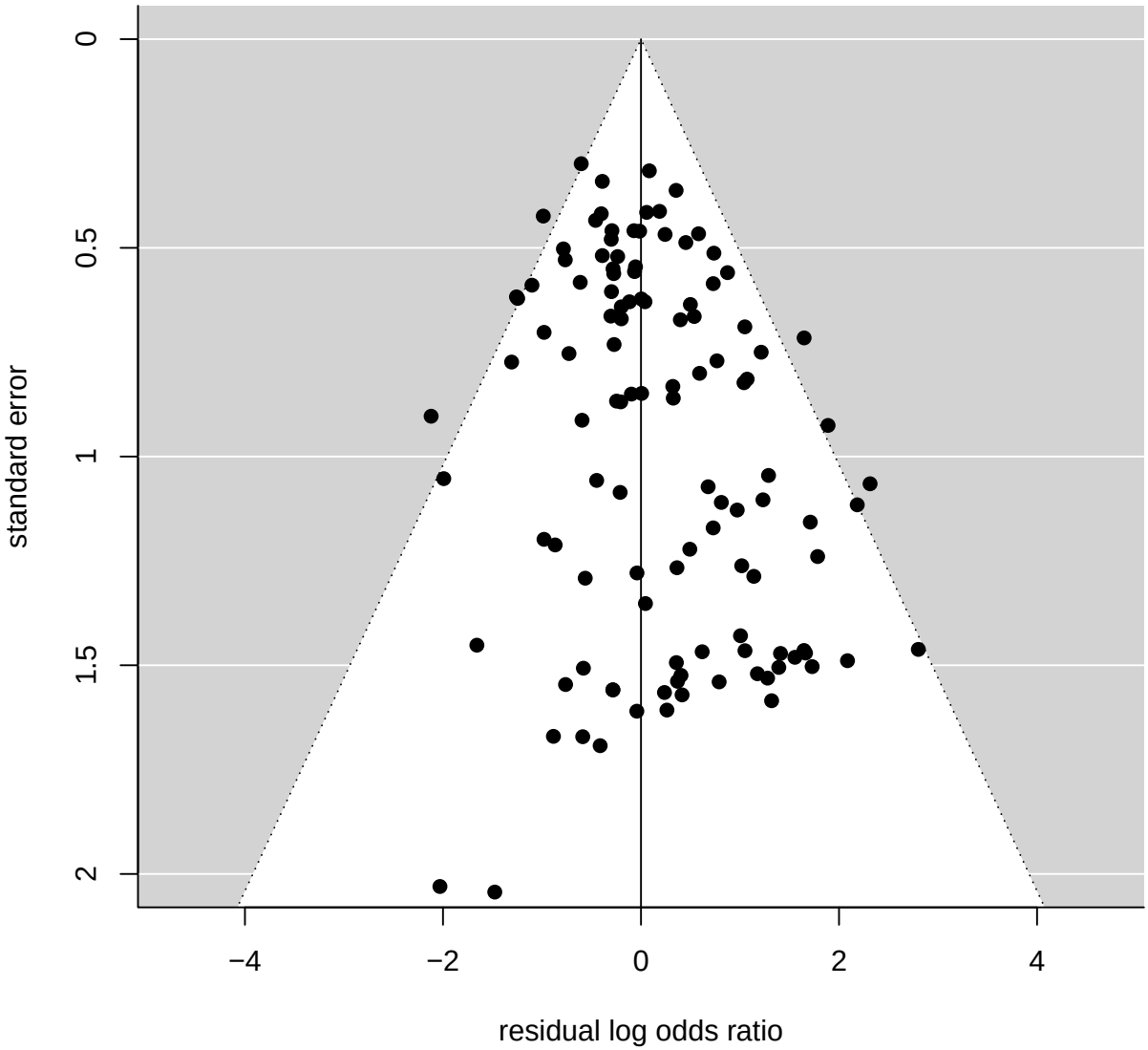

Figure 5: Funnel plot of model residuals: For each observed log odds ratio, the x-axis gives the residuals from a univariate mixed-effects model that includes the same study characteristics as the bivariate model. The y-axis gives the standard errors of the observed log odds ratios, and the triangle defines the 95 percent confidence interval. In the absence of publication bias, the funnel plot is symmetric; asymmetry indicates publication bias. We can observe that the plot appears largely symmetric with slightly more studies having positive residuals than negative ones.

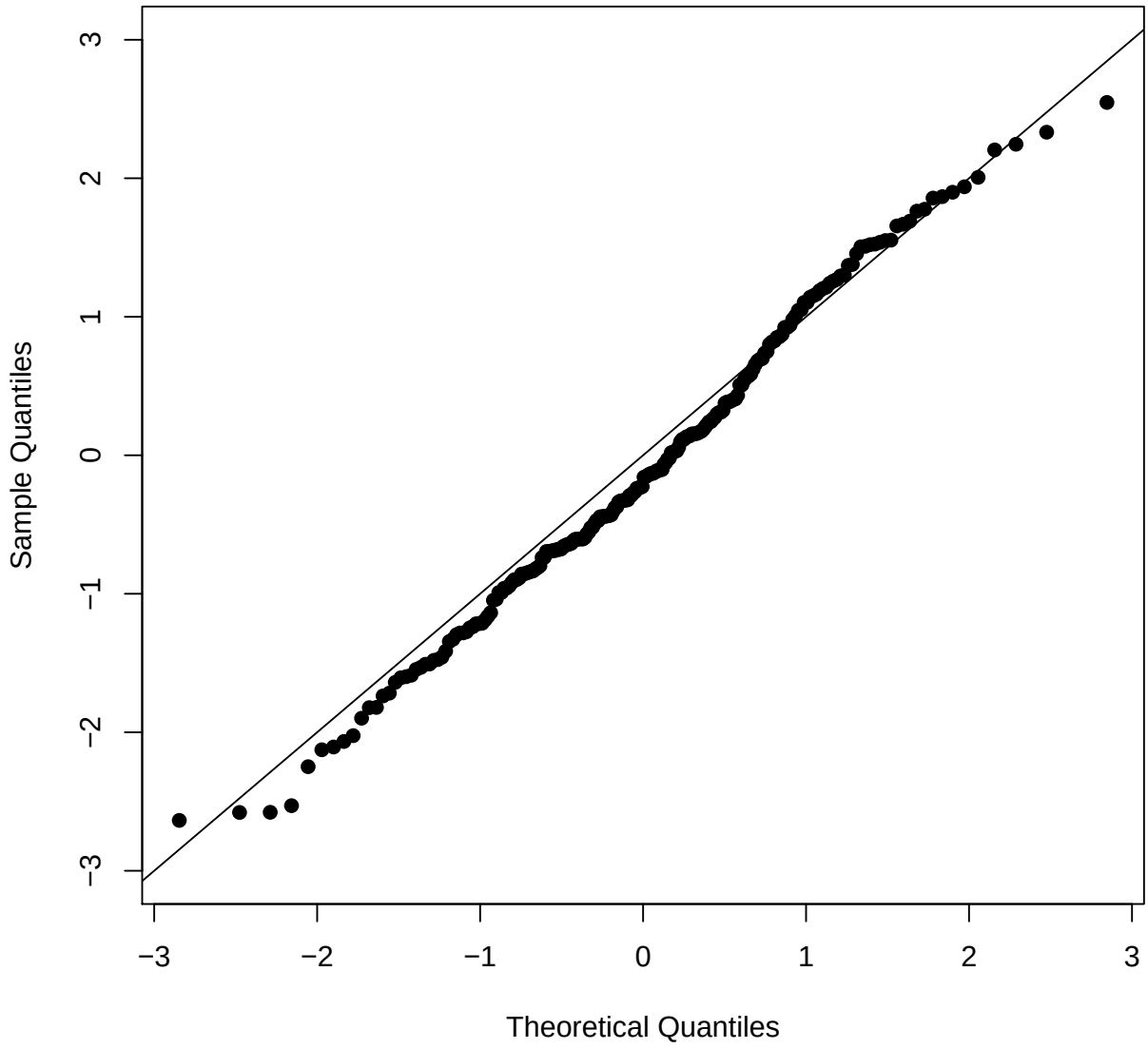

Figure 6: Quantile Plot of Residuals: When the residuals follow a normal distribution, they should line up along the diagonal line where the theoretical quantiles of a normal distribution equal the observed quantiles. Indeed, the observed quantiles follow approximately a normal distribution and deviations are small and non-systematic.

Table 1
Summary of Study Characteristics

		counts	
Name	Description	no	yes
short menu	short menu used in both formats	199	27
three hypotheses	problem complicated by introducing third hypothesis	208	18
two or more cues	problem complicated by introducing additional cues or cue values	212	14
probability question	probability question asked in natural frequency format	206	20
frequency question	frequency question asked in conditional probability format	216	10
enumerated population	enumerated population given in conditional probability format	198	28
multiple events	conditional probability format phrased in terms of multiple events	162	64
visual aid	visual aid used in both formats	200	26
base rate*	base rate in problem solved by participants	numerical	
hit rate*	hit rate in problem solved by participants	numerical	
false-alarm rate*	false-alarm rate in problem solved by participants	numerical	
show-up fee*	participants are paid a show-up fee	50	115
performance pay*	participants are paid a by performance	153	12
strict scoring	correct answer based on accuracy and/or protocol	193	33
within-subject*	within-subjects study design	212	14
both formats	both formats solved by each participant	174	52
additional problems	number of additional problems solved by each participant	numerical	
high numeracy*	participants have a high numeracy scores	26	26
experts	educated participants (students & professionals)	38	188

Notes: Study characteristics marked with * were coded but not able to be included in the full-sample meta-analysis because of the lack of a sufficient number of logits; where needed, we conducted separate subset-analyses that did not control for the larger set of study characteristics.

Table 2
Estimated Average Proportions and Implied Odds Ratios

		natural frequency		conditional probability			
No	variable	Δ_m	$\text{CI}_{.95}$	Δ_m	$\text{CI}_{.95}$	OR	$\text{CI}_{.95}$
0	baseline	.24	[.13, .40]	.04	[.01, .14]	7.11	[4.37, 11.56]
1	short menu	+.12	[−.13, +.47]	+.11	[−.00, +.38]	3.08	[1.75, 5.42]
2	three hypotheses	+.19	[−.05, +.45]	+.09	[+.01, +.25]	4.66	[2.32, 9.37]
3	two or more cues	−.04	[−.20, +.37]	−.02	[−.04, +.07]	11.37	[4.59, 28.20]
4	probability question	−.02	[−.17, +.27]	+.02	[−.02, +.13]	4.42	[2.08, 9.37]
5	frequency question	−.01	[−.18, +.38]	−.02	[−.04, +.06]	12.40	[3.90, 39.44]
6	enumerated population	−.00	[−.15, +.24]	−.00	[−.02, +.06]	7.06	[3.40, 14.69]
7	multiple events	+.02	[−.17, +.37]	+.02	[−.02, +.12]	5.53	[3.31, 9.24]
8	visual aid	+.23	[+.02, +.45]	+.22	[+.04, +.53]	2.52	[1.36, 4.67]
9	strict scoring	−.02	[−.16, +.23]	+.01	[−.03, +.15]	4.72	[2.67, 8.33]
10	both formats	+.13	[−.15, +.53]	−.01	[−.03, +.09]	16.19	[8.75, 29.96]
11	additional problems	+.01	[−.03, +.05]	+.00	[−.01, +.02]	6.72	[4.14, 10.91]
12	experts	+.07	[−.11, +.34]	+.03	[−.03, +.24]	5.96	[4.39, 8.09]
0	chances	.40	[.16, .71]	.19	[.03, .67]	2.77	[1.49, 5.16]
0	no incentive	.41	[.27, .57]	.10	[.05, .21]	6.10	[3.46, 10.77]
1	show-up fee	−.03	[−.22, +.20]	+.01	[−.07, +.21]	4.90	[3.60, 6.68]
2	performance pay	+.23	[+.07, +.36]	+.11	[−.00, +.28]	6.62	[2.57, 17.07]
0	low numeracy	.26	[.00, .97]	.04	[.00, .56]	9.37	[4.10, 21.41]
1	high numeracy	+.25	[+.04, +.46]	+.11	[−.01, +.50]	5.92	[3.32, 10.57]

Notes: Top panel gives results of full-sample meta-analysis and bottom panels give results of subset analyses; numbers without sign give baseline proportions and numbers with sign give changes in proportions; confidence intervals based on cluster-robust standard errors (Hedges et al., 2010); numbers are rounded.

Table 3
Data of Meta-Analysis

Estimate	format	c	i	short menu	2+ cues	3 hypot.	prob. quest.	freq. quest.	enum. pop.	mult. events	vis. aid	show fee	perf. pay	strict scor.	both formats	high num.	exp.	add. probs.
Gigerenzer and Hoffrage (1995)																		
A) exp 1: standard	F	14	16	N	N	N	N	N	N	Y	N	Y	N	Y	Y	—	Y	14
A) exp 1: standard	P	5	25	N	N	N	N	N	N	Y	N	Y	N	Y	Y	—	Y	14
B) exp 1: short	F	15	15	Y	N	N	N	N	N	Y	N	Y	N	Y	Y	—	Y	14
B) exp 1: short	P	8	22	Y	N	N	N	N	N	Y	N	Y	N	Y	Y	—	Y	14
Hoffrage and Gigerenzer (1998)																		
A) ⌀ all four problems	F	11	13	N	N	N	N	N	N	Y	N	—	—	Y	Y	—	Y	1
A) ⌀ all four problems	P	2	22	N	N	N	N	N	N	Y	N	—	—	Y	Y	—	Y	1
Krauss et al. (1999)																		
A) version 1 vs 3	F	22	19	N	Y	N	N	N	N	Y	N	—	—	N	N	—	Y	0
A) version 1 vs 3	P	5	36	N	Y	N	N	N	N	Y	N	—	—	N	N	—	Y	0
Mellers and McGraw (1999)																		
A) exp 1: Nat, standard	F	13	33	N	N	N	N	N	N	Y	N	N	N	N	N	—	Y	0
A) exp 1: CP, standard	P	3	39	N	N	N	N	N	N	Y	N	N	N	N	N	—	Y	0
B) exp 1: Nat, joint	F	9	33	Y	N	N	N	N	N	Y	N	N	N	N	N	—	Y	0
B) exp 1: CP, joint	P	1	38	Y	N	N	N	N	N	Y	N	N	N	N	N	—	Y	0
C) exp 1: Sys, standard	P	7	28	N	N	N	N	Y	Y	N	N	N	N	N	N	—	Y	0
D) exp 1: Sys, joint	P	7	37	Y	N	N	N	Y	Y	N	N	N	N	N	N	—	Y	0
E) exp 2: Nat, standard	F	10	121	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
E) exp 2: CP, standard	P	5	117	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
F) exp 2: Nat, joint	F	24	115	Y	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
F) exp 2: CP, joint	P	15	81	Y	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
G) exp 2: Sys, standard	P	7	131	N	N	N	N	Y	Y	N	N	Y	N	N	N	—	Y	0
Evans et al. (2000)																		
A) exp 3: NF easy	F	8	20	N	N	N	Y	N	N	N	N	—	—	N	N	—	Y	7
A) exp 3: CP hard	P	5	21	N	N	N	Y	N	N	N	N	—	—	N	N	—	Y	7
Fiedler et al. (2000)																		
A) exp 1: incompatible	F	14	35	N	N	N	Y	N	Y	N	N	Y	N	N	N	—	Y	3
A) exp 1: incompatible	P	9	36	N	N	N	Y	N	Y	N	N	Y	N	N	N	—	Y	3
B) exp 1: common	F	24	29	Y	N	N	Y	N	Y	N	N	Y	N	N	N	—	Y	3
B) exp 1: common	P	20	31	Y	N	N	Y	N	Y	N	N	Y	N	N	N	—	Y	3
Hoffrage, Lindsey, Hertwig, and Gigerenzer (2000)																		
A) example 1, apps in medicine	F	43	53	N	N	N	N	N	N	Y	N	Y	N	N	Y	—	Y	0
A) example 1, apps in medicine	P	13	83	N	N	N	N	N	N	Y	N	Y	N	N	Y	—	Y	0
Macchi (2000)																		
A) PF	F	12	18	N	N	N	N	N	Y	Y	N	—	—	Y	N	—	Y	0
A) NPP	P	2	28	N	N	N	N	N	Y	Y	N	—	—	Y	N	—	Y	0
B) NPF	P	1	29	N	N	N	N	Y	Y	N	N	—	—	Y	N	—	Y	0
Sedlmeier and Gigerenzer (2001)																		
A) exp 2: strict, trees, test2	F	18	6	N	N	N	Y	N	N	Y	Y	Y	N	N	N	—	Y	6
A) exp 2: strict, trees, test2	P	18	6	N	N	N	Y	N	N	Y	Y	Y	N	N	N	—	Y	6
Lindsey et al. (2003)																		
A) law students, p(profile)	F	51	76	N	N	N	Y	N	Y	Y	N	Y	N	N	Y	—	Y	0
A) law students, p(profile)	P	1	126	N	N	N	Y	N	Y	Y	N	Y	N	N	Y	—	Y	0
B) jurists, p(profile)	F	20	7	N	N	N	Y	N	Y	Y	N	N	N	N	Y	—	Y	0
B) jurists, p(profile)	P	3	24	N	N	N	Y	N	Y	Y	N	N	N	N	Y	—	Y	0
Yamagishi (2003)																		
A) exp 1: viz	F	34	13	N	N	Y	N	N	N	Y	Y	Y	N	N	N	—	Y	0
A) exp 1: viz	P	30	13	N	N	Y	N	N	N	Y	Y	Y	N	N	N	—	Y	0
B) exp 1: no viz	F	17	23	N	N	Y	N	N	N	Y	N	Y	N	N	N	—	Y	0
B) exp 1: no viz	P	7	31	N	N	Y	N	N	N	Y	N	Y	N	N	N	—	Y	0
C) exp 2: viz	F	33	10	N	N	Y	N	N	N	Y	Y	Y	N	N	N	—	Y	0
C) exp 2: viz	P	34	8	N	N	Y	N	N	N	Y	Y	Y	N	N	N	—	Y	0
D) exp 2: no viz	F	25	20	N	N	Y	N	N	N	Y	N	Y	N	N	N	—	Y	0
D) exp 2: no viz	P	5	36	N	N	Y	N	N	N	Y	N	Y	N	N	N	—	Y	0
E) exp 3: viz a	F	160	42	N	N	Y	N	N	N	Y	Y	Y	N	N	N	—	Y	0
E) exp 3: viz a	P	157	40	N	N	Y	N	N	N	Y	Y	Y	N	N	N	—	Y	0
F) exp 3: viz b	F	141	60	N	N	Y	N	N	N	Y	Y	Y	N	N	N	—	Y	0

Notes: c and i denote the number of correct and incorrect responses, respectively; when $c = 0$ or $i = 0$, 0.5 was added to all counts of the same experiment; Y denotes the presence and N denotes the absence of a study characteristic; effects from studies including chances formats (Brase, 2008; Girotto & Gonzalez, 2001; Sirota, Kostovičová, & Vallée-Tourangeau, 2015a) are excluded; table includes only the moderators used in the analyses.

Table 3
Data of Meta-Analysis (continued)

Estimate	format	c	i	short menu	2+ cues	3 hypot.	prob. quest.	freq. quest.	enum. pop.	mult. events	vis. aid	show fee	perf. pay	strict scor.	both formats	high num.	exp.	add. probs.
F) exp 3: viz b	P	119	79	N	N	Y	N	N	N	Y	Y	Y	N	N	N	—	Y	0
Hoffrage and Gigerenzer (2004)																		
A) med students, short menu	F	63	33	Y	N	N	N	N	N	Y	N	Y	N	N	Y	—	Y	0
A) med students, short menu	P	39	57	Y	N	N	N	N	N	Y	N	Y	N	N	Y	—	Y	0
O'Brien et al. (2004)																		
A) exp 2: curta	F	12	13	Y	N	N	N	Y	N	Y	N	—	—	N	N	—	Y	0
A) exp 2: curta	P	3	22	Y	N	N	N	Y	N	Y	N	—	—	N	N	—	Y	0
B) exp 2: padrao	F	3.5	22.5	N	N	N	N	Y	N	Y	N	—	—	N	N	—	Y	0
B) exp 2: padrao	P	.5	25.5	N	N	N	N	Y	N	Y	N	—	—	N	N	—	Y	0
Bramwell et al. (2006)																		
A) pregnant women	F	3	18	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0
A) pregnant women	P	1	21	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0
B) companions	F	3	17	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0
B) companions	P	3	17	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0
C) midwives	F	.5	20.5	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0
C) midwives	P	.5	22.5	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0
D) obstetricians	F	13	7	N	N	N	N	N	N	Y	N	N	N	N	N	—	Y	0
D) obstetricians	P	1	20	N	N	N	N	N	N	Y	N	N	N	N	N	—	Y	0
Zhu and Gigerenzer (2006)																		
A) exp 1: adults	F	17	6	N	N	N	N	N	N	Y	N	—	—	N	N	—	Y	6
A) exp 1: adults	P	11	12	N	N	N	N	N	N	Y	N	—	—	N	N	—	Y	6
B) exp 2: adults	F	23	7	N	N	N	N	N	N	Y	N	—	—	N	N	—	Y	9
B) exp 2: adults	P	17	13	N	N	N	N	N	N	Y	N	—	—	N	N	—	Y	9
C) exp 2: 4th grade	F	6.5	24.5	N	N	N	N	N	N	Y	N	—	—	N	N	—	N	9
C) exp 2: 4th grade	P	.5	10.5	N	N	N	N	N	N	Y	N	—	—	N	N	—	N	9
D) exp 2: 5th grade	F	13.5	17.5	N	N	N	N	N	N	Y	N	—	—	N	N	—	N	9
D) exp 2: 5th grade	P	.5	10.5	N	N	N	N	N	N	Y	N	—	—	N	N	—	N	9
E) exp 2: 6th grade	F	14.5	16.5	N	N	N	N	N	N	Y	N	—	—	N	N	—	N	9
E) exp 2: 6th grade	P	.5	10.5	N	N	N	N	N	N	Y	N	—	—	N	N	—	N	9
Brase (2008)																		
B) exp 1: NF	F	9.5	19.5	N	N	N	N	Y	N	Y	N	Y	N	N	N	—	Y	0
B) exp 1: Normalized	P	.5	28.5	N	N	N	N	Y	N	Y	N	Y	N	N	N	—	Y	0
Konheim-Kalkstein (2008)																		
A) exp 1: correct solutions	F	7	24	N	N	N	Y	N	N	Y	N	—	—	N	N	—	Y	0
A) exp 1: correct solutions	P	2	29	N	N	N	Y	N	N	Y	N	—	—	N	N	—	Y	0
B) exp 4: lo num	F	2	16	N	N	N	Y	N	Y	N	N	—	—	N	N	N	Y	0
B) exp 4: lo num	P	1	23	N	N	N	Y	N	Y	N	N	—	—	N	N	N	Y	0
C) exp 4: hi num	F	9	26	N	N	N	Y	N	Y	N	N	—	—	N	N	Y	Y	0
C) exp 4: hi num	P	8	23	N	N	N	Y	N	Y	N	N	—	—	N	N	Y	Y	0
Brase (2009a)																		
A) course req., no viz	F	12.5	39.5	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
A) course req., no viz	P	.5	50.5	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
B) course req., viz	F	21	31	N	N	N	N	N	Y	Y	Y	Y	N	N	N	—	Y	0
B) course req., viz	P	14	36	N	N	N	N	N	Y	Y	Y	Y	N	N	N	—	Y	0
C) flat pay, no viz	F	14	34	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
C) flat pay, no viz	P	2	45	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
D) flat pay, viz	F	16	32	N	N	N	N	N	Y	Y	Y	Y	N	N	N	—	Y	0
D) flat pay, viz	P	13	36	N	N	N	N	N	Y	Y	Y	Y	N	N	N	—	Y	0
E) var. pay, no viz	F	23	19	N	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
E) var. pay, no viz	P	4	38	N	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
F) var. pay, viz	F	28	14	N	N	N	N	N	Y	Y	Y	N	Y	N	N	—	Y	0
F) var. pay, viz	P	17	25	N	N	N	N	N	Y	Y	Y	N	Y	N	N	—	Y	0
Chapman and Liu (2009)																		
A) medical, lo num	F	5.5	60.5	N	N	N	N	N	N	Y	N	Y	N	N	Y	N	Y	0
A) medical, lo num	P	.5	87.5	N	N	N	N	N	N	Y	N	Y	N	N	Y	N	Y	0
B) car, lo num	F	9.5	78.5	N	N	N	N	N	N	N	N	Y	N	N	Y	N	Y	0
B) car, lo num	P	.5	65.5	N	N	N	N	N	N	N	N	Y	N	N	Y	N	Y	0
C) medical, hi num	F	26	66	N	N	N	N	N	N	Y	N	Y	N	N	Y	Y	Y	0
C) medical, hi num	P	1	91	N	N	N	N	N	N	Y	N	Y	N	N	Y	Y	Y	0
D) car, hi num	F	29	63	N	N	N	N	N	N	N	N	Y	N	N	Y	Y	Y	0

Notes: c and i denote the number of correct and incorrect responses, respectively; when $c = 0$ or $i = 0$, 0.5 was added to all counts of the same experiment; Y denotes the presence and N denotes the absence of a study characteristic; effects from studies including chances formats (Brase, 2008; Girotto & Gonzalez, 2001; Sirota, Kostovičová, & Vallée-Tourangeau, 2015a) are excluded; table includes only the moderators used in the analyses.

Table 3
Data of Meta-Analysis (continued)

Estimate	format	c	i	short menu	2+ cues	3 hypot.	prob. quest.	freq. quest.	enum. pop.	mult. events	vis. aid	show fee	perf. pay	strict scor.	both formats	high num.	exp.	add. probs.
D) car, hi num	P	7	85	N	N	N	N	N	N	N	N	Y	N	N	Y	Y	Y	0
Galesic, Gigerenzer, and Straubinger (2009)																		
A) older adults, hi num	F	3	10	N	N	N	N	N	N	Y	N	Y	N	N	N	Y	N	1
A) older adults, hi num	P	2	7	N	N	N	N	N	N	Y	N	Y	N	N	N	Y	N	1
B) younger adults, hi num	F	15	28	N	N	N	N	N	N	Y	N	Y	N	N	N	Y	Y	1
B) younger adults, hi num	P	4	38	N	N	N	N	N	N	Y	N	Y	N	N	N	Y	Y	1
C) older adults, lo num	F	2.5	9.5	N	N	N	N	N	N	Y	N	Y	N	N	N	N	N	1
C) older adults, lo num	P	.5	14.5	N	N	N	N	N	N	Y	N	Y	N	N	N	N	N	1
D) younger adults, lo num	F	1.5	12.5	N	N	N	N	N	N	Y	N	Y	N	N	N	N	Y	1
D) younger adults, lo num	P	.5	17.5	N	N	N	N	N	N	Y	N	Y	N	N	N	N	Y	1
Misuraca et al. (2009)																		
A) ø cond. 1-4 vs 5	F	27.5	93.5	N	N	N	N	N	N	Y	N	N	N	N	N	—	Y	0
A) ø cond. 1-4 vs 5	P	.5	30.5	N	N	N	N	N	N	Y	N	N	N	N	N	—	Y	0
Moro et al. (2011)																		
A) exp 1: NF	F	7.5	14.5	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
A) exp 1: Nested-sets CP	P	.5	21.5	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
D) exp 2: gemstone, viz	F	13	5	N	N	N	N	N	N	Y	Y	Y	N	N	Y	—	Y	0
D) exp 2: gemstone, viz	P	1	20	N	N	N	N	N	N	Y	Y	Y	N	N	Y	—	Y	0
E) exp 2: gemstone, no viz	F	16	2	N	N	N	N	N	N	Y	N	Y	N	N	Y	—	Y	0
E) exp 2: gemstone, no viz	P	1	22	N	N	N	N	N	N	Y	N	Y	N	N	Y	—	Y	0
Siegrist and Keller (2011)																		
A) exp 1: mammography	F	15	117	N	N	N	N	N	N	Y	N	—	—	Y	N	—	N	0
A) exp 1: mammography	P	1	133	N	N	N	N	N	N	Y	N	—	—	Y	N	—	N	0
B) exp 2: mammography	F	9	68	N	N	N	N	N	N	Y	N	—	—	Y	N	—	N	0
B) exp 2: mammography	P	1	70	N	N	N	N	N	N	Y	N	—	—	Y	N	—	N	0
C) exp 3: ø social & cookie	F	55	81	N	N	N	N	N	N	Y	N	—	—	Y	N	—	N	1
C) exp 3: ø social & cookie	P	9	61	N	N	N	N	N	N	Y	N	—	—	Y	N	—	N	1
Tsai, Miller, and Kirlik (2011)																		
A) NF	F	6	6	N	N	N	N	N	N	Y	N	Y	N	Y	N	—	Y	5
A) CP	P	4	8	N	N	N	N	N	N	Y	N	Y	N	Y	N	—	Y	5
Hill and Brase (2012)																		
A) exp 2: ø med & New B., lo num	F	1.5	32.5	N	N	N	N	N	N	Y	N	Y	N	N	N	N	Y	1
A) exp 2: ø med & New B., lo num	P	.5	37.5	N	N	N	N	N	N	Y	N	Y	N	N	N	N	Y	1
B) exp 2: ø med & New B., hi num	F	10	36	N	N	N	N	N	N	Y	N	Y	N	N	N	Y	Y	1
B) exp 2: ø med & New B., hi num	P	2	45	N	N	N	N	N	N	Y	N	Y	N	N	N	Y	Y	1
C) exp 3: ø med & New B., lo num	F	1	49	N	N	N	N	N	N	Y	N	Y	N	N	N	N	Y	1
C) exp 3: ø med & New B., lo num	P	1	58	N	N	N	N	N	N	Y	N	Y	N	N	N	N	Y	1
D) exp 3: ø med & New B., hi num	F	9	38	N	N	N	N	N	N	Y	N	Y	N	N	N	Y	Y	1
D) exp 3: ø med & New B., hi num	P	3	38	N	N	N	N	N	N	Y	N	Y	N	N	N	Y	Y	1
Ferguson and Starmer (2013)																		
A) experts, incentive, short	F	8	3	Y	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
A) experts, incentive, short	P	8	5	Y	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
B) novices, incentive, short	F	15	7	Y	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
B) novices, incentive, short	P	4	19	Y	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
C) experts, incent., standard	F	9	3	N	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
C) experts, incent., standard	P	1	7	N	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
D) novices, incent., standard	F	11.5	14.5	N	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
D) novices, incent., standard	P	.5	24.5	N	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
E) experts, no incent., short	F	8	4	Y	N	N	N	N	N	Y	N	—	—	N	N	—	Y	0
E) experts, no incent., short	P	1	9	Y	N	N	N	N	N	Y	N	—	—	N	N	—	Y	0
F) novices, no incent., short	F	16	18	Y	N	N	N	N	N	Y	N	—	—	N	N	—	Y	0
F) novices, no incent., short	P	5	28	Y	N	N	N	N	N	Y	N	—	—	N	N	—	Y	0
G) experts, no incent., stand.	F	6	6	N	N	N	N	N	N	Y	N	—	—	N	N	—	Y	0
G) experts, no incent., stand.	P	1	11	N	N	N	N	N	N	Y	N	—	—	N	N	—	Y	0
H) novices, no incent., stand.	F	11	24	N	N	N	N	N	N	Y	N	—	—	N	N	—	Y	0
H) novices, no incent., stand.	P	1	33	N	N	N	N	N	N	Y	N	—	—	N	N	—	Y	0
Johnson and Tubau (2013)																		
A) exp 1: complicated, hi num	F	20	6	N	N	N	N	N	N	N	N	Y	N	N	Y	Y	Y	0
A) exp 1: complicated, hi num	P	2	24	N	N	N	N	N	N	N	N	Y	N	N	Y	Y	Y	0
B) exp 1: complicated, lo num	F	6.5	16.5	N	N	N	N	N	N	N	N	Y	N	N	Y	N	Y	0

Notes: c and i denote the number of correct and incorrect responses, respectively; when $c = 0$ or $i = 0$, 0.5 was added to all counts of the same experiment; Y denotes the presence and N denotes the absence of a study characteristic; effects from studies including chances formats (Brase, 2008; Girotto & Gonzalez, 2001; Sirota, Kostovičová, & Vallée-Tourangeau, 2015a) are excluded; table includes only the moderators used in the analyses.

Table 3
Data of Meta-Analysis (continued)

Estimate	format	c	i	short menu	2+ cues	3 hypot.	prob. quest.	freq. quest.	enum. pop.	mult. events	vis. aid	show fee	perf. pay	strict scor.	both formats	high num.	exp.	add. probs.
B) exp 1: complicated, lo num	P	.5	22.5	N	N	N	N	N	N	N	N	Y	N	N	Y	N	Y	0
C) exp 1: simple, hi num	F	30	4	N	N	Y	N	N	N	N	N	Y	N	N	Y	Y	Y	0
C) exp 1: simple, hi num	P	12	22	N	N	Y	N	N	N	N	N	Y	N	N	Y	Y	Y	0
D) exp 1: simple, lo num	F	12	3	N	N	Y	N	N	N	N	N	Y	N	N	Y	N	Y	0
D) exp 1: simple, lo num	P	1	14	N	N	Y	N	N	N	N	N	Y	N	N	Y	N	Y	0
E) exp 2: long, hi num	F	16	5	N	N	N	N	N	N	N	N	Y	N	N	Y	Y	Y	0
E) exp 2: long, hi num	P	4	16	N	N	N	N	N	N	N	N	Y	N	N	Y	Y	Y	0
F) exp 2: long, lo num	F	9.5	11.5	N	N	N	N	N	N	N	N	Y	N	N	Y	N	Y	0
F) exp 2: long, lo num	P	.5	23.5	N	N	N	N	N	N	N	N	Y	N	N	Y	N	Y	0
G) exp 2: short, hi num	F	18	2	N	N	N	N	N	N	N	N	Y	N	N	Y	Y	Y	0
G) exp 2: short, hi num	P	5	16	N	N	N	N	N	N	N	N	Y	N	N	Y	Y	Y	0
H) exp 2: short, lo num	F	17.5	6.5	N	N	N	N	N	N	N	N	Y	N	N	Y	N	Y	0
H) exp 2: short, lo num	P	.5	20.5	N	N	N	N	N	N	N	N	Y	N	N	Y	N	Y	0
Lesage et al. (2013)																		
A) exp 1: total sample, relative	F	18.5	23.5	N	N	N	N	N	Y	N	N	Y	N	N	N	—	Y	1
A) exp 1: total sample, relative	P	.5	42.5	N	N	N	N	N	Y	N	N	Y	N	N	N	—	Y	1
B) exp 1: no tot.samp., relative	F	17	26	N	N	N	N	N	N	N	N	Y	N	N	N	—	Y	1
B) exp 1: no tot.samp., relative	P	2	41	N	N	N	N	N	N	N	N	Y	N	N	N	—	Y	1
C) exp 1: total sample, absolute	F	18	26	Y	N	N	N	N	Y	N	N	Y	N	N	N	—	Y	1
C) exp 1: total sample, absolute	P	15	29	Y	N	N	N	N	Y	N	N	Y	N	N	N	—	Y	1
D) exp 1: no tot.samp., absolute	F	27	14	Y	N	N	N	N	N	N	N	Y	N	N	N	—	Y	1
D) exp 1: no tot.samp., absolute	P	13	30	Y	N	N	N	N	N	N	N	Y	N	N	N	—	Y	1
E) exp 2: relative	F	18	24	N	N	N	N	N	N	N	N	N	N	N	N	—	N	1
E) exp 2: relative	P	6	44	N	N	N	N	N	N	N	N	N	N	N	N	—	N	1
F) exp 2: absolute	F	30	21	Y	N	N	N	N	N	N	N	N	N	N	N	—	N	1
F) exp 2: absolute	P	21	26	Y	N	N	N	N	N	N	N	N	N	N	N	—	N	1
Friederichs et al. (2014)																		
A) NF, no viz	F	11	18	N	N	N	Y	N	N	Y	N	—	—	N	N	—	Y	2
A) CP, no viz	P	5	29	N	N	N	Y	N	N	Y	N	—	—	N	N	—	Y	2
Sirota, Juanchich, and Hagmayer (2014)																		
A) exp 2: med & children's	F	86	65	N	N	N	N	N	N	Y	N	—	—	Y	Y	—	Y	3
A) exp 2: med & children's	P	30	121	N	N	N	N	N	N	Y	N	—	—	Y	Y	—	Y	3
Binder et al. (2015)																		
A) mammography, no viz, task 1	F	1.5	19.5	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0
A) mammography, no viz, task 1	P	.5	22.5	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0
B) mammography, viz tree, task 1	F	8.5	15.5	N	N	N	N	N	N	Y	Y	N	N	N	N	—	N	0
B) mammography, viz tree, task 1	P	.5	20.5	N	N	N	N	N	N	Y	Y	N	N	N	N	—	N	0
C) economics, no viz, task 1	F	11	10	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0
C) economics, no viz, task 1	P	1	20	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0
D) economics, viz tree, task 1	F	11	9	N	N	N	N	N	N	Y	Y	N	N	N	N	—	N	0
D) economics, viz tree, task 1	P	1	20	N	N	N	N	N	N	Y	Y	N	N	N	N	—	N	0
Hill and Brase (2015)																		
A) MTurk, ST	F	9.5	31.5	N	N	N	N	N	N	N	N	Y	N	N	N	—	N	2
A) MTurk, ST	P	.5	37.5	N	N	N	N	N	N	N	N	Y	N	N	N	—	N	2
B) MTurk, ø BC & CF	F	10	30	N	Y	N	N	N	N	N	N	Y	N	N	N	—	N	2
B) MTurk, ø BC & CF	P	1	36	N	Y	N	N	N	N	N	N	Y	N	N	N	—	N	2
C) online, ST	F	1.5	45.5	N	N	N	N	N	N	N	N	Y	N	N	N	—	Y	2
C) online, ST	P	.5	46.5	N	N	N	N	N	N	N	N	Y	N	N	N	—	Y	2
D) online, ø BC & CF	F	4	42	N	Y	N	N	N	N	N	N	Y	N	N	N	—	Y	2
D) online, ø BC & CF	P	1	45	N	Y	N	N	N	N	N	N	Y	N	N	N	—	Y	2
E) paper, ST	F	7	42	N	N	N	N	N	N	N	N	Y	N	N	N	—	Y	2
E) paper, ST	P	1	47	N	N	N	N	N	N	N	N	Y	N	N	N	—	Y	2
F) paper, ø BC & CF	F	5.5	44.5	N	Y	N	N	N	N	N	N	Y	N	N	N	—	Y	2
F) paper, ø BC & CF	P	.5	48.5	N	Y	N	N	N	N	N	N	Y	N	N	N	—	Y	2
Hoffrage, Krauss, et al. (2015)																		
A) exp 1: two hyp, three cue values	F	18.5	14.5	N	Y	N	N	N	N	Y	N	—	—	Y	Y	—	Y	1
A) exp 1: two hyp, three cue values	P	.5	32.5	N	Y	N	N	N	N	Y	N	—	—	Y	Y	—	Y	1
B) exp 1: three hyp, one dichot cue	F	12	20	N	N	Y	N	N	N	Y	N	—	—	Y	Y	—	Y	1
B) exp 1: three hyp, one dichot cue	P	3	29	N	N	Y	N	N	N	Y	N	—	—	Y	Y	—	Y	1
C) exp 1: two hyp, two dichot cues	F	11.5	21.5	N	Y	N	N	N	N	Y	N	—	—	Y	Y	—	Y	1
C) exp 1: two hyp, two dichot cues	P	.5	32.5	N	Y	N	N	N	N	Y	N	—	—	Y	Y	—	Y	1

Notes: c and i denote the number of correct and incorrect responses, respectively; when $c = 0$ or $i = 0$, 0.5 was added to all counts of the same experiment; Y denotes the presence and N denotes the absence of a study characteristic; effects from studies including chances formats (Brase, 2008; Girotto & Gonzalez, 2001; Sirota, Kostovičová, & Vallée-Tourangeau, 2015a) are excluded; table includes only the moderators used in the analyses.

Table 3
Data of Meta-Analysis (continued)

Estimate	format	c	i	short menu	2+ cues	3 hypot.	prob. quest.	freq. quest.	enum. pop.	mult. events	vis. aid	show fee	perf. pay	strict scor.	both formats	high num.	exp.	add. probs.
D) exp 1: two hyp, three dichot cues	F	12	20	N	Y	N	N	N	N	Y	N	—	—	Y	Y	—	Y	1
D) exp 1: two hyp, three dichot cues	P	1	31	N	Y	N	N	N	N	Y	N	—	—	Y	Y	—	Y	1
Hoffrage, Hafenbrädl, and Bouquet (2015)																		
A) undergraduates, all tasks	F	53	79	N	N	N	N	N	N	N	N	—	—	Y	N	—	Y	1
A) undergraduates, all tasks	P	19	108	N	N	N	N	N	N	N	N	—	—	Y	N	—	Y	1
B) junior executives, all tasks	F	23	34	N	N	N	N	N	N	N	N	—	—	Y	N	—	Y	1
B) junior executives, all tasks	P	18	45	N	N	N	N	N	N	N	N	—	—	Y	N	—	Y	1
C) senior executives, all tasks	F	11	19	N	N	N	N	N	N	N	N	—	—	Y	N	—	Y	1
C) senior executives, all tasks	P	7	24	N	N	N	N	N	N	N	N	—	—	Y	N	—	Y	1
Vallée-Tourangeau, Abadie, and Vallée-Tourangeau (2015)																		
A) exp 1&2: lo interact., hi num	F	17	11	N	N	N	N	N	N	Y	N	N	N	N	N	Y	Y	2
A) exp 1&2: lo interact., hi num	P	3	24	N	N	N	N	N	N	Y	N	N	N	N	N	Y	Y	2
B) exp 1&2: hi interact., hi num	F	20	6	N	N	N	N	N	N	Y	Y	N	N	N	N	Y	Y	2
B) exp 1&2: hi interact., hi num	P	21	12	N	N	N	N	N	N	Y	Y	N	N	N	N	Y	Y	2
C) exp 1&2: lo interact., lo num	F	7	10	N	N	N	N	N	N	Y	N	N	N	N	N	N	Y	2
C) exp 1&2: lo interact., lo num	P	1	17	N	N	N	N	N	N	Y	N	N	N	N	N	N	Y	2
D) exp 1&2: hi interact., lo num	F	12	7	N	N	N	N	N	N	Y	Y	N	N	N	N	N	Y	2
D) exp 1&2: hi interact., lo num	P	4	8	N	N	N	N	N	N	Y	Y	N	N	N	N	N	Y	2
E) exp 3: standard	F	7.5	10.5	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
E) exp 3: standard	P	.5	17.5	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
F) exp 3: fleshed out	F	5.5	13.5	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
F) exp 3: fleshed out	P	.5	18.5	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0

Notes: c and i denote the number of correct and incorrect responses, respectively; when $c = 0$ or $i = 0$, 0.5 was added to all counts of the same experiment; Y denotes the presence and N denotes the absence of a study characteristic; effects from studies including chances formats (Brase, 2008; Girotto & Gonzalez, 2001; Sirota, Kostovičová, & Vallée-Tourangeau, 2015a) are excluded; table includes only the moderators used in the analyses.