Memory & Cognition (2021) 49:235–248
https://doi.org/10.3758/s13421-020-01087-5

Which cognitive individual differences predict good Bayesian reasoning? Concurrent comparisons of underlying abilities

Gary Brase[1]

Published online: 19 August 2020

Abstract
We know a lot about how to present Bayesian reasoning tasks in order to aid performance, but less about underlying individual differences that can account for interindividual variability on the same tasks. Such information would be useful for both theoretical and practical reasons. Two theoretical positions, ecological rationality and nested set views, generate multiple hypotheses about which individual difference traits should be most relevant as underlying Bayesian reasoning performance. However, because many of these traits are somewhat overlapping, testing variables in isolation can yield misleading results. The present research assesses Bayesian reasoning abilities in conjunction with multiple individual different measures. Across three experiments, Bayesian reasoning was best predicted by measures of numerical literacy and visuospatial ability, as opposed to several different measures of cognitive thinking dispositions/styles, ability to conceptually model set-theoretic relationships, or cognitive processing ability (working memory span). These results support an ecological rationality view of Bayesian reasoning, rather than nested sets views. There also was some predictive ability for the Cognitive Reflection Task, which was only partially due to the numeracy aspects of that instrument, and further work is needed to clarify if this is a distinct factor. We are now beginning to understand not only how to build Bayesian reasoning tasks, but also how to build good Bayesian reasoners.

Keywords Bayesian reasoning · Individual differences · Numerical literacy · Spatial ability · Ecological rationality · Nested sets

Introduction

An array of practical situations involve Bayesian reasoning, which involves revising the baseline likelihood of an event happening, given new information. Examples include interpreting how a medical test result (new information) effects the prior likelihood someone has a disease, using new evidence to update (from some baseline estimate) how likely a defendant is to be guilty, and recognizing how recent interactions may have influenced a relationship partner's previous level of satisfaction. Research on Bayesian reasoning abilities, however, have traditionally painted a distressing picture; Casscells et al. (1978) for instance found that only 18% of doctors and medical students were able to correctly use Bayesian reasoning to interpret a medical test result.

Subsequent research has established several ways to present Bayesian reasoning tasks in order to generally improve performance. Two of these key methods are giving the relevant information as whole numbers embedded within a natural sampling framework (i.e., natural frequencies) and including a pictorial representation of the information (see McDowell & Jacobs, 2017, for a meta-analysis). The facilitating effects of these task presentation formats have been demonstrated across legal, medical, and business settings (e.g., Garcia-Retamero & Hoffrage, 2013; Gigerenzer et al., 2007; Hoffrage & Gigerenzer, 1998; Hoffrage et al., 2000, 2015; Lindsey et al., 2003).

The present research focusses not on characteristics of Bayesian reasoning tasks but rather on the characteristics of the persons doing the reasoning. In particular, *what are the individual differences that best predict Bayesian reasoning performance?* This question is important because, although presentation format guides are undoubtedly useful (e.g., for presentations to patients, jurors, or stockholders), it is also important to know which presenters and receivers of

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13421-020-01087-5) contains supplementary material, which is available to authorized users.

✉ Gary Brase
gbrase@ksu.edu

[1] Department of Psychological Sciences, Kansas State University, North Manhattan, KS, USA

information have the underlying dispositions and abilities to understand those presentations.

Research on individual differences associated with Bayesian reasoning ability has largely centered on numerical literacy (the ability to understand and work with numbers, often shortened to "numeracy"; e.g., Chapman & Liu, 2009), but that focus is widening now to include individual differences in other cognitive abilities. Another specific trait that appears to predict Bayesian reasoning is visuospatial ability (Brase & Hill, 2017; Kellen et al., 2013). Other studies have looked at more general constructs associated with Bayesian reasoning such as cognitive resources (Lesage et al., 2013), cognitive ability (Stanovich & West, 2000), cognitive reflection (Sirota & Juanchich, 2011), cognitive processing (Sirota et al., 2014), and working memory (Yin et al., 2020). A subsequent section develops how the diversity of these accounts, and the broad generality of their proposed predictors, present practical and theoretical challenges.

Numeracy and visuospatial ability

It makes some sense that differences in numeracy are associated with Bayesian reasoning because it is, after all, a numerically based task. There has been some disagreement, though, regarding how and why the two are related. Similar disagreement has been emerging about the interpretation of other individual differences predicting Bayesian reasoning (i.e., visuospatial ability, general cognitive abilities or resources), and the issues tend to align with two different theoretical viewpoints.

One long-standing general theoretical viewpoint about Bayesian reasoning is the *ecological rationality approach (ERA)* view, which stresses that people should perform better on tasks (including Bayesian reasoning) when they are presented in ways that are more consistent with the natural and evolutionary ecology of the real-world environment. This view stresses considerations of the mind being evolved to expect certain types of information (Brase et al., 1998) and mental processes being "fast and frugal"; taking advantage of structural regularities and intercorrelations that occur in the environment generally (whether evolutionarily recurrent or modern; Gigerenzer & Hoffrage, 1995). The ERA interprets the facts that Bayesian reasoning improves when tasks are given using naturally sampled frequencies (e.g., Gigerenzer & Hoffrage, 1995) and pictorial representation (e.g., Garcia-Retamero et al., 2010) as being due to those presentations better mapping onto the ecology of the natural environment (i.e., a world that predominately includes visually individuated, countable items that can be organized as they are encountered).

An ERA view suggests that measures of numeracy and visuospatial ability contribute fairly directly to Bayesian reasoning performances, as measures of how well individuals can transfer the written and drawn Bayesian tasks into an ecologically more suitable representation. Consistent with this, numeracy has been found to predict Bayesian reasoning performance across both levels of numeracy (e.g., high vs. low numeracy; Hill & Brase, 2012) and different measures of numerical literacy (Brase & Hill, 2017; Johnson & Tubau, 2013). Brase and Hill (2017) also found that multiple measures of visuospatial ability were correlated with Bayesian reasoning, and that this relationship (when dichotomized) was consistent across levels of ability.

Another long-standing theoretical viewpoint is the *nested sets approach* (NSA) view, which stresses the general ability to perceive the nested sets structure of any given Bayesian reasoning task and tends to generally eschew evolutionary and ecological rationality explanations. The NSA view is that both naturally sampled frequencies and pictures improve the "clarity" or help people to "see" the nested set structure of the Bayesian reasoning task. This has also variously been described as inducing a "partative formulation" (e.g., Macchi, 2000) as facilitating the "construction of a set inclusion mental model" (Evans et al., 2000), and as making nested relationships "opaque" or "transparent" (Sloman et al., 2003).

An NSA view suggests that numeracy predicts Bayesian reasoning performance indirectly, via more general abilities (e.g., being more able to see nested sets structures or even broader cognitive abilities; e.g., Sirota & Juanchich, 2011; Sirota et al., 2014). Better visuospatial ability, like the use of pictures to aid Bayesian reasoning, also helps by allowing people to be better able to understand the nested sets structure. Overall, the NSA view leads to the prediction that any variable that increases the general ability to form a good nested set task representation will therefore be associated with better Bayesian reasoning. That also implies that those variables will consistently tend to interact with each other, such that their predictive power will be diminished whenever other transparency-enhancing variables are also included as factors. In other words, if there is one general underlying factor (nested sets transparency) that is facilitated in multiple ways by specific manipulations (use of pictures, frequencies, higher numeracy, higher visuospatial ability, etc.), then the measured effectiveness of each of those manipulations will initially be strong but then appear to diminish as additional concurrent manipulations are assessed.

Different NSA views

The NSA view has more recently been incorporated by some advocates within either a mental-models framework (Johnson-Laird et al., 1999; Pighin et al., 2015) or a dual-systems framework (e.g., Barbey & Sloman, 2007; Lesage et al., 2013). Different versions of the NSA view lead to

various different expectations about which individual differences will be more central and predictive of Bayesian reasoning performance.

NSA as a dual-system disposition Some NSA views have adopted a dual-systems model. There are some variations (see Evans & Stanovich, 2013, for one overview), but dual-systems models generally propose that people's thinking occurs in one of two broad modes: System 1 and System 2 processes. System 1 processes tend to be more implicit, intuitive, and automatic; a faster and less effortful. System 2 processes tend to be more explicit, conscious, and controlled; deliberative and effortful. According to this view, frequencies, natural sampling, and pictures all help promote a better understanding of the nested sets relationship inherent in Bayesian reasoning, and thereby invoke System 2 processes to correctly reason about the problem (e.g., Barbey & Sloman, 2007; Evans et al., 2000; Sirota, Kostovičová, & Vallée-Tourangeau, 2015; Sloman et al., 2003).

One dual-systems model of the NSA view then is that people who have a disposition to employ System 2 processes will thereby be better able to analyze and reason about Bayesian reasoning tasks. This is consistent, for example, with Sloman et al. (2003, p. 298), which uses the terminology of "inside" versus "outside" views:

> "This [representing instances can reveal the set structure of a problem] is a fairly direct consequence of representing the instances of the categories in a problem. These instances make up the sets or classes that correspond (in an outside view) to the categories. Most representational schemes that identify instances and the categories they belong to will automatically also specify the set structures relating the categories."

According to this version of the NSA view, a disposition toward greater task engagement and effort would be a strong driver of better performance because it promotes clearer or deeper thinking about the problem. This appears to correspond well to "need for cognition," a long-standing and highly validated trait construct of people's dispositions to deeply engage with cognitively difficult tasks. People high in need for cognition enjoy thinking about topics as they are presented, enjoy the process of thinking, and readily apply their thinking skills in effective ways (Cacioppo & Petty, 1982, 1984).

NSA as a dual-system ability Another possible dual-systems model of the NSA view is that better *abilities* to access and use System 2 cognitive processes (rather than disposition to do so) should improve Bayesian reasoning (see Stanovich & West, 1998a, 1998b, 2000, 2008, for a more general approach along these lines). The exact nature of these general abilities is not well defined, but Sirota et al. (2014) claimed that two measures (the Cognitive Reflection Task and the Raven advanced progressive matrices) were appropriate and significantly predicted Bayesian reasoning performances.

The Cognitive Reflection Task (CRT; Frederick, 2005) has actually been found to predict Bayesian performance a few times (Lesage et al., 2013; Sirota & Juanchich, 2011). Sirota et al. (2014), however, employed it as a measure of "general reasoning," and argued that this relationship indicated "the involvement of a general reasoning mechanism as postulated by the NSA, rather than the involvement of a specialized cognitive mechanism operating automatically, as posited by the ERA" (p. 201). There is considerable debate, however, about what the CRT actually measures, and it is known to partially overlap with numeracy (e.g., Campitelli & Gerrans, 2014; Thompson & Oppenheimer, 2016). Subsequent work (Sirota et al., 2018) has recognized that the CRT theoretically confounds mathematical ability (i.e., numeracy) with whatever else it measures and that it can be psychometrically problematic. Adding to this ambiguity, the analyses in Sirota et al. (2014) used a median split of the three-item CRT scores, which can be very problematic (e.g., Irwin & McClelland, 2003; MacCallum et al., 2002).

The Raven advanced progressive matrices (Raven et al., 1977) as a measure of general cognitive abilities was also a significant predictor of Bayesian reasoning performance in Sirota et al. (2014). This test also has a problematic theoretical confound, though. It consists of a series of visual geometric designs, each with a missing piece, and the test-taker is asked to pick from a set of options to complete the missing piece. There are ongoing debates about the degree to which the Raven's results are a measure of just general cognitive ability, or if it is also assessing visuospatial ability (e.g., Colom et al., 2004; DeShon et al., 1995; Vigneau & Bors, 2005). Thus, it is again possible that the Sirota et al. (2014) results are due to a confound, this time with visuospatial ability.

Part of the challenge for this view, clearly, is how to accurately measure "general ability" given that it is an ambiguous construct. Evans and Stanovich (2013) have suggested that working memory is a defining feature of System 2 processes, but that was not assessed in Sirota et al. (2014). In any event, a good measure of working memory (as a general cognitive-processing ability) will need to not be theoretically confounded with numeracy or visuospatial ability.

NSA as a style of mental model The NSA view has also been placed within the general framework of mental models (e.g., Johnson-Laird et al., 1999; Johnson-Laird et al., 2015). The mental-models framework explains the efficacy of format manipulations (frequencies, natural sampling, pictures, etc.) in improving Bayesian reasoning in terms of more effective construction of a mental model of the reasoning task.

As with the dual-systems framework, though, it is not always clear whether this is due to one's *disposition* or one's *ability* to construct mental models. One possibility is to construe greater task engagement as the driver of better performance, in much the same way as in the dual-systems framework. This interpretation is consistent with the third "fundamental principle" of mental-models theory: "with deliberation reasoners can use the meaning of assertions to flesh out mental models into fully explicit models." (e.g., Johnson-Laird et al., 2015, p. 204). If this is the case, the Need for Cognition Scale is a viable option for assessing individual differences in the tendency to put more effort into fully fleshing out mental models.

NSA as a model ability Another possibility for framing the NSA view within the mental-models theory is to maintain that there is a more specific ability for mental model construction of nested sets. Johnson-Laird (1983) claims multiple representation formats are possible within mental models, including tokens, spatial relations, and temporal or causal relations (p. 410). Johnson-Laird goes on to further delineate between "physical" models and "conceptual" models, with six major types of physical models and four major types of conceptual models (monadic, relational, meta-linguistic, and set-theoretic models). Apparently, then, the idea of nested set mental models as the underlying foundation for effective Bayesian reasoning is not just any mental model but a rather specific type: *conceptual set-theoretic models*. Indeed, this interpretation is most consistent with the mental-models view of the past 20 years (e.g., Johnson-Laird et al., 1999, 2015).

It follows, then, that individual differences in ability to construct conceptual set-theoretic mental models should be a better predictor of reasoning performance than other possible predictors. A difficulty arises at this point, though, because there are no known psychological measures of individual differences in abilities to understand and reason specifically with set-theoretic models. Part of the present research, therefore, includes the development of assessments designed to measure nested sets modeling abilities.

Hypotheses

All viewpoints reviewed here recognize that there are individual differences that should be meaningfully related to Bayesian reasoning abilities and performance.[1] The key differences are in which individual differences are primary factors in accounting for variations in Bayesian reasoning abilities. Crucially, some of these measures are likely to be intercorrelated to some extent (e.g., Brase & Hill, 2017), so looking at individual variables in isolation would risk finding significant correlations driven by unstudied third variables (e.g., see prior discussion of results in Sirota et al., 2014). The current research therefore concurrently measures multiple individual difference traits alongside the target behavior of Bayesian reasoning.

In summary, hypotheses about the relationship between various individual differences measures and Bayesian reasoning performance differ according to which theoretical view is entertained and which version of the view is considered:

- The *ERA* predicts that both numeracy and visuospatial ability will best predict Bayesian reasoning performance. The predictive power of these will not be strongly affected by other individual differences, such as thinking dispositions, the ability to use nested sets, or general cognitive processing abilities.
- The *NSA as thinking disposition/mental model style views* predict that one's tendency to more extensively think about tasks (either more System 2 processing or fleshing out of one's model) will best predict Bayesian reasoning performance. These views imply that the predictive power of thinking disposition or style (e.g., need for cognition) would dominate those other variables.
- An *NSA as set-theoretic modeling ability view* predicts that the ability to construct conceptual set-theoretic models will best predict Bayesian reasoning performance, and that this relationship will not be significantly diminished by other variables.
- An *NSA as System 2 ability view* predicts that general processing ability (e.g., working memory) will significantly predict Bayesian reasoning performance, and that its predictive power will largely subsume the predictive ability of other traits.

It does not appear that the later three hypotheses (all versions of the NSA view) are required to be mutually exclusive, even as they are distinct. While that is possibly problematic in terms of generating specific and testable predictions for the general NSA view, it also opens up the present opportunity to work out the more specific versions of this view. Additionally, by using concurrent assessments of multiple relevant individual difference measures it may be possible to identify even overlapping but supportive outcomes.

[1] Sirota, Juanchich, and Hagmayer (2014) argue that the ecological rationality view effectively predicts "no cognitive involvement" (p. 198) and that "individual differences in cognitive processing should not predict Bayesian performance with natural frequencies, since processing should result automatically" (p.199). No citations are given for this particular prediction, and it is not clear where this radical interpretation came from.

Experiment 1

Methods

Participants

Data were collected from 352 undergraduates at a large public university. The relatively large sample size was chosen to ensure sufficient power, given the number of individual difference variables included in this research and analyses. (With nine predictor variables, a .05 alpha level, .80 power, and a small effect size (f^2 = .05), G*Power (Faul et al., 2009) recommends a sample size of 322.) For their participation, students received credit toward partial fulfillment of their introductory psychology course or extra credit in an upper-level psychology course. After eliminating 48 people who did not complete the entire study, the remaining 304 participants were used for analyses. The average age of participants was 19.96 years (SD = 2.21), and 89 of the participants (29.3%) were male.

Materials and procedures

The protocol for this and all subsequent experiments was reviewed and declared exempt by the relevant Institutional Review Board. The research was implemented using an online survey platform (Qualtrics), and the elements of the study included three main sections. After reading information about the study and giving informed consent to participate, the first section consisted of demographic questions and the revised version of the Need for Cognition Scale (NCS; Cacioppo, Petty, & Kao, 1984). The NCS consists of 18 items, presented in random order, that measure how much people enjoy the process and act of thinking about topics. An example NCS item is, "I would prefer complex to simple problems." Each item was rated on a 1–5 scale, and an overall score for each individual was a sum of all the ratings. The NCS showed very good reliability (Cronbach's α =.872), and the present sample had a mean of 58.45 and standard deviation of 11.04.

The second section consisted of four Bayesian reasoning tasks that covered a range of presentation styles and contents, but all used a natural sampling organization for the numerical information. The differences across tasks were in the use of a pictorial aid or not, in the whole numbers describing a single event (chances of an event) or multiple events (frequency of events), and in the background story of each task. Half of the tasks included pictures, half were in each numerical presentation type, and all four tasks were given in random order. Details about these tasks contents, variations, and in-depth analyses of their relative performances are given in Brase (in press) and the stimuli are available at https://osf.io/cb5yh/?view_only=b253f3f251884eae89134270f3c2aee0. For the present research, the number of Bayesian reasoning tasks correctly answered (i.e., from 0 to 4) was used as an overall measure of Bayesian reasoning ability, and participants' mean overall score was 1.78 (*SD* = 1.37).

The third section consisted of eight measures, presented in random order and with randomized item orders within each measure. Three of these were measures of numerical literacy, three of these were measures of visuospatial ability, and two were newly developed measures of nested sets thinking. The three numerical literacy measures were from Brase and Hill (2017):

a) A multiple-choice version of the general numeracy scale (MCQ-GNS; Hill et al., 2019), which consists of 11 items, each with the most common four answers as alternatives. For example, one item in this scale is, "Which of the following represents the biggest risk of getting a disease?" (0.8%, 1%, 5%, or 10%), and an overall score for each individual was a tally of correct answer choices. The MCQ-GNS showed good reliability (Cronbach's α = .762), and the present sample had a mean of 8.18 and standard deviation of 2.53.
b) The three-item Berlin Numeracy Test (Cokely et al., 2012), which did not have very good reliability (Cronbach's α =.465). This low reliability is likely because this test has only three items, which are designed specifically to bisect samples. The present sample had a mean of 0.57 and standard deviation of 0.80.
c) The Expanded Cognitive Reflection Task (ECRT; Toplak et al., 2014), which consists of seven items (three from the original CRT and four new items) that all involve quantitative thinking and suppressing misleading intuitive answers. For example, one item in this scale is "Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are in the class?" The ECRT showed good reliability (Cronbach's α =.742), and the present sample had a mean of 4.12 and standard deviation of 2.03.

The three measures of visuospatial ability were also from Brase and Hill (2017):

a) The Paper Folding Test (Ekstrom et al., 1976) consists of 20 items, each of which contains a series of drawings of a uniquely folded square piece of paper with a hole punched through the folded paper in a specific location. For each drawing, participants are then asked to judge what the folded paper would look like, once unfolded, by selecting one of five pictures of unfolded papers located to the right of the original drawings. The Paper Folding Test showed very good reliability (Cronbach's α =.880), and the present sample had a mean of 14.24 and standard deviation of 6.43.

b) The Water Level Task consists of seven items that are a computer-based adaptation of the Piagetian conservation task (Brase & Hill, 2017). Each item asked participants to evaluate three drawings of a tilted glass, half filled with water (signified by a blue line). The three glasses were all tilted at the same angle, but the water line differed in each image. The participant is asked to pick which picture is closest to the correct appearance of an actual glass with water, being held at that angle (see Online Supplemental Materials in Brase & Hill, 2017, for full stimuli). The Water Level Task showed only weak reliability (Cronbach's α =.639), and the present sample had a mean of 4.05 and standard deviation of 1.93.

c) The Mental Rotation of Figures Test (Peters et al., 1995; Peters & Battista, 2008) included 24 items that each presented a target figure and asked participants to select which two out of four other drawings were rotated versions of the target figure. The Mental Rotation Test showed excellent reliability (Cronbach's α = .918), and the present sample had a mean of 34.09 and standard deviation of 9.01.

Two measures were created to assess abilities to think and reason about sets (the nested sets as set-theoretic modeling ability hypothesis):

a) The Set Theory Test (STT) was a test of ability to perform basic operations in set theory. Twelve items tested a person's ability to reason about unions, intersections, and differences of number sets (the symbols and meanings of these operations were provided to participants, and every item page included a reminder of what each symbol meant). Four of the items used just two sets (A = {*1, 2, 3*} and B = {*1, 2, 4, 5*}), asking the four possible set operations given these two sets. The other eight items used four sets (A = {*2, 3, 4, 5*}; B = {*4, 5, 6, 7*}; C = {*6, 7, 8, 9*}; D = {*8, 9, 10, 11*}) and asked about various set operations utilizing two of the four sets. The STT (and the EST described below) is given in its entirety at https://osf.io/cb5yh/?view_only=b253f3f251884eae89134270f3c2aee0. It showed excellent reliability (Cronbach's α = .925), and the present sample had a mean of 5.17 and standard deviation of 4.29.

b) The Euler Syllogism Test (EST) was based on a similar logic as the test of ability to generate alternative representations (Torrens et al., 1999). The six items of the EST each show three intersecting Euler circles that were labeled with letters. Participants were asked to select which of four categorical conclusions (e.g., "All B are also C") was a correct description of the above picture. The EST initially showed very weak reliability (Cronbach's α = .576), but further analysis indicated that by removing Item 4 the reliability became acceptable (Cronbach's α = .745). After doing so, the present sample had a mean of 4.51 and standard deviation of 1.04.

Results

The data for all experiments are available at https://osf.io/cb5yh/?view_only=b253f3f251884eae89134270f3c2aee0, along with supplemental analyses. None of the experiments were preregistered. Before proceeding to the focal analyses, correlations between the different individual difference measures were evaluated. Need for Cognition was correlated with most of the other measures (r = .084-.331), the three measures of visuospatial ability were intercorrelated (r = .413-.692, all $p < .001$), the three measures of numeracy were intercorrelated (r = .361-.516, all $p < .001$), and the two measures of nested sets ability were correlated (r = .247, $p < .001$; see analyses in the Online Supplemental Material). Also, however, there were correlations across these different types of measures (r = .161-.577, all $p < .01$). Although these correlations are not problematically high, they led us to specifically evaluate possible multicollinearity issues in subsequent analyses.

First, a simultaneous entry multiple regression was conducted with all the potential predictor variables. This model was statistically significant ($F(9, 294) = 19.155$, $p < .001$), with an overall adjusted R^2 of .350. Just three measures were significant predictors, though: The Paper Folding Task (p = .031), the general numeracy scale (p = .002), and the Cognitive Reflection Test (p = .016). The Need for Cognition scale was not a significant predictor, but is worth noting as possibly having some utility (p = .070). Looking at the variance inflation factor (VIF) as an indication of multicollinearity, the VIFs were all between 1.22 and 2.34. These values (i.e., between 1 and 5) indicate moderate correlations between predictors, but that they are not problematically multicollinear (see analyses in the Online Supplemental Material for full simultaneous entry regression details).

Following up this model, two separate hierarchical multiple regressions were run, based on the differing predictions of the theoretical views outlined earlier. Both analyses entered Need for Cognition first (assessing this as a predictor to possibly weaken all further predictors, based on the cognitive disposition/style view). Need for Cognition, indeed, was a significant predictor when entered first (adjusted R^2 of .092; Table 1, see full results in analyses in the Online Supplemental Material).

At the second level either the visuospatial and numeracy measures were entered or the nested-sets modeling measures were entered (Table 1). Based on the NSA as a modeling ability view, adding the nested sets thinking measures (after

Table 1 Summary of results from hierarchical multiple regression analyses of Bayesian reasoning task performance, entering potential predictors at different levels per hypotheses

	R	R^2	Adjusted R^2
Numeracy and visuospatial ability first			
1: (Constant), Need for Cognition	.308	.095	**.092**
2: Model 1 + Berlin Numeracy, Water Level, General Numeracy, Mental Rotation, CRT, Paper Folding	.604	.364	**.349**
3: Model 2 + Euler Circles, Sets Reasoning	.608	.370	**.350**
Nested sets ability first			
1: (Constant), Need for Cognition	.308	.095	**.092**
2: Model 1 + Euler Circles, Sets Reasoning	.482	.232	**.225**
3: Model 2 + Water Level, Berlin Numeracy, General Numeracy, CRT, Mental Rotation, Paper Folding	.608	.370	**.350**

need for cognition) should supersede all the other potential predictors, whereas adding the numeracy and visuospatial predictors first should not subsume the nested sets predictors. On the other hand, the ERA predicts that adding the numeracy and visuospatial predictors first should supersede the nested-sets modeling measures whereas adding the nested-sets modeling measures first should not subsume the numeracy and visuospatial predictors.

Table 1 shows that when the numeracy and visuospatial predictors were entered as the second step, the adjusted R^2 significantly increased to .349, but the variance accounted for was not improved at the third step as nested-sets modeling measures were added (adjusted R^2 of .350). When the nested-sets modeling measures were entered as the second step (Table 2), the overall adjusted R^2 significantly increased to .225 and then significantly increased again (to .350) with the addition of numeracy and visuospatial predictors at the third step. Both models, of course, end with all the same predictors, converging with the simultaneous entry model results at the third step.

These results indicate that numeracy and visuospatial abilities more fundamentally underlie individual differences in Bayesian reasoning performances, whereas thinking disposition/style and nested sets modeling abilities are not significant underlying factors other than to the extent they overlap with other abilities. Such a result clearly supports the ERA rather than the NSA as a thinking disposition/style or mental-modelling ability view.

Another way to appreciate the present results is to look at performances of participants at both the high and the low ends of these predictive ability measures. When considering just participants who were half a standard deviation above the mean on all three significant predictors (Paper Folding Task, General Numeracy Scale, and Cognitive Reflection Test), those 43 people gave correct Bayesian responses 75.6% of the time. At the other extreme, participants half a standard deviation below the mean on all three significant predictors (n = 41) gave correct Bayesian responses 12.2% of the time.

Experiment 2

The results of Experiment 1 seem to provide evidence in support of the ERA, as opposed to a few possible NSA views. However, some methodological factors could possibly have influenced those results, and Experiment 2 was designed to address these. First, there could have been some idiosyncratic combination of context stories and presentation formats

Table 2 Summary of results from hierarchical multiple regression analyses of Bayesian reasoning task performance, entering potential predictors at different levels per hypotheses

	R	R^2	Adjusted R^2
Numeracy and visuospatial ability first			
1: (Constant), REI-Rationality, REI-Experientiality	.363	.131	**.120**
2: Model 1 + General Numeracy, Paper Folding	.565	.319	**.300**
3: Model 2 + Euler Circles, Sets Reasoning, Sets Reas. 2	.568	.323	**.289**
Nested sets ability first			
1: (Constant), REI-Rationality, REI-Experientiality	.363	.131	**.120**
2: Model 1 + Euler Circles, Sets Reasoning, Sets Reas. 2	.477	.228	**.201**
3: Model 2 + General Numeracy, Paper Folding	.568	.323	**.289**

within the Bayesian reasoning problems that influenced performance and thus influenced the subsequent analyses (see Brase, in press). Second, one could argue there are alternatives for assessing cognitive style other than the Need for Cognition scale. For example, Sirota et al. (2014) tested the effect of cognitive abilities and thinking dispositions using different – albeit problematically confounded – measures. The Rational/Experiential Inventory (REI-40; Pacini & Epstein, 1999) was also used by Sirota et al. (2014) and may work better than Need for Cognition as a general thinking style measure.

Third, Experiment 1 included two new measures designed to assess ability to mentally model and reason with sets, and these need further evaluation. In particular, although these measures are high in face validity there could be a concern that the Set Theory Test involved working with sets of numbers (between 1 and 11) and therefore might tap into some aspect of numeracy. The Euler Syllogism Test had some reliability issues in Experiment 1, so Experiment 2 again evaluated this measure's reliability. Additionally, Experiment 2 also added a third measure designed to assess ability to model and reason about sets, the Sets Theory Test II, which uses almost no numbers and uses the picture-based format of the Euler Syllogisms Test but without the formal logic aspect.

Finally, it could be argued that the hierarchical regression models in Experiment 1 were loaded in favor of the ERA hypotheses simply because there were three potential predictor variables for numeracy and three potential predictor variables for visuospatial ability, but just two measures of nested sets thinking (one with low reliability). The present experiment was therefore designed to be methodologically loaded in the opposite direction regarding the number of potential predictors.

Methods

Participants

Data were collected from 177 undergraduates at a large public university, using the same recruitment process as in Experiment 1. The sample size was reduced, compared to Experiment 1, based on the strength of those results and the fewer possible predictor variables. After eliminating people who did not complete the entire study (n = 21) and finished the entire study in less than 5 min (n = 7), the remaining 149 participants were used for analyses. The average age of participants was 19.52 years (SD = 5.71), and 64 of the participants (43.0%) were male.

Materials and procedures

The procedure of this experiment was the same as Experiment 1, but with different configurations of materials in the three sections. The first section of the study consisted of a few demographic questions and the Rational/Experiential Inventory (REI-40; Pacini & Epstein, 1999). The REI-40 consists of 40 items that measure rational and experiential processing as two unipolar dimensions. The rational thinking scale is designed to assess logical and analytical thinking, and the experiential thinking scale is designed to assess the use of feelings and intuitions in making decisions. For example, one item in the rationality scale is "I have a logical mind," whereas an item in the experiential scale is "I believe in trusting my hunches." Each item was rated from 1 to 5 for self-descriptiveness and showed very good reliability for both the Rationality scale (Cronbach's α = .830) and the Experientiality scale (.864). The Rationality scale had a mean of 6.89 (standard deviation of 0.99), whereas the Experientiality scale had a mean of 6.93 (standard deviation of 1.00). Consistent with prior work (Pacini & Epstein, 1999), the two scales were not correlated with each other (r =-.076, p = .359).

The second section consisted of four Bayesian reasoning tasks, with the same variations as described in Experiment 1, but the context stories of these tasks were systematically reversed in both numerical presentation and picture presence. An overall measure of Bayesian reasoning performance (from 0 to 4) provided a general measure of Bayesian reasoning ability. Participants' mean overall score on the Bayesian reasoning tasks was 1.64 (SD = 1.31).

The third section consisted of five measures, presented in random order and with randomized item orders within each measure. One of these was the multiple-choice numeracy scale (MCQ-GNS; Hill, Brase, & Kenney, 2019), which showed acceptable reliability (Cronbach's α = .716) and had a mean of 8.41 and a standard deviation of 2.33. Another measure (of visuospatial ability) was the Paper Folding Test (Ekstrom et al., 1976), which again showed very good reliability (Cronbach's α = .880) and had a mean of 11.26 and a standard deviation of 5.12. The three other measures all assessed abilities to think and reason about sets, including the Set Theory Test (STT; Cronbach's α = .912, mean of 4.60 and standard deviation of 4.05) and the Euler Syllogism Test (EST). As in Experiment 1, the EST had poor initial reliability (Cronbach's α = .408), but after removing Item 4 the reliability was close to acceptable (Cronbach's α = .637, mean of 4.42, and standard deviation of 1.016). The third measure of ability to model and reason about sets, the Set Theory Test II (STT2), was developed for the present study. In contrast to the SST, the SST2 uses a combination of letters and numbers as the items in the sets and gives participants five multiple-choice options to select from as answers. The SST2 furthermore uses a graphic representation of sets as circles, labeled with Greek symbols, to provide the sets information (see https://osf.io/cb5yh/?view_only=b253f3f251884eae89134270f3c2aee0 for the full text of these tasks). The STT2 showed very

good reliability (Cronbach's α = .882), and the present sample had a mean of 4.87 and standard deviation of 1.87.

Results

Performance across the individual difference measures were significantly intercorrelated to a large extent, with the exception of the REI Experientiality subscale (see analyses in the Online Supplemental Material for a full table of these results). Because of this, the subsequent analyses evaluated possible multicollinearity issues.

First, a simultaneous entry multiple regression was conducted with all the potential predictor variables included. This model was statistically significant ($F(7,141) = 9.766$, $p < .001$), with an overall adjusted R^2 of .293. Just two measures were significant predictors, though: visuospatial ability (Paper Folding Task, $p < .001$) and numeracy ($p = .027$). The variance inflation factors (VIF) were all between 1.02 and 1.81, indicating no problematic multicollinearity (see analyses in the Online Supplemental Material for full simultaneous entry regression details).

Following this, two separate hierarchical multiple regressions were run, based on the differing predictions of the theoretical perspectives outlined earlier. Both analyses entered the REI subscales first as a general thinking disposition/style measure. REI was a significant predictor when entered first (adjusted R^2 of .120; $F(2,146) = 11.050$, $p < .001$), with the Rationality subscale driving this result ($p < .001$).

At the second level either the visuospatial and numeracy measures were entered or the three nested sets thinking measures were entered (Table 2). When the numeracy and visuospatial predictors were entered as the second step, the adjusted R^2 significantly increased to .300. At this step the REI no longer was a significant predictor, replaced by visuospatial ability ($p < .001$) and numeracy ($p = .002$). The variance accounted for was not improved at the third step with the addition of nested sets ability predictors (adjusted R^2 of .293) and the significant predictors remained visuospatial ability ($p < .001$) and numeracy ($p = .027$).

When the nested sets predictors were entered as the second step (Table 2), the overall adjusted R^2 significantly increased to .212, driven by both the Rationality subscale of the REI ($p = .002$) and the Euler Reasoning Scale ($p = .027$). At the third step, however, both these predictors fell out of statistical significance with the addition of numeracy and visuospatial predictors, and the variance accounted for significantly increased again (to .293). Both models, of course, converged with the simultaneous entry model results at the third step. The full regression results are provided in the analyses in the Online Supplemental Material.

As with Experiment 1, another way to understand the present results is to compare participants at both the high and the low ends of these predictive individual difference ability measures. When considering just participants who were half a standard deviation above the mean on both significant predictors (visuospatial ability and numeracy), those 33 people gave correct Bayesian responses 66.7% of the time. At the other extreme, participants half a standard deviation below the mean on both significant predictors (n = 25) gave correct Bayesian responses 19.0% of the time.

Experiment 3

Experiment 2 replicated the results of Experiment 1, using a different measure of general cognitive disposition/style, systematic re-arranging of Bayesian reasoning task context stories, and a set of potential predictor variables numerically loaded to favor NSA views. Bayesian reasoning performance was significantly predicted by thinking style and ability initially (the Rational/Experiential Inventory), but it was better accounted for by numeracy and visuospatial ability. Subset reasoning ability, however, did not add to that predictive ability.

Neither Experiment 1 nor Experiment 2, though, assessed the hypothesis that the NSA view, as part of a dual-systems framework, is based on ability to deploy System 2 abilities. By that interpretation, individual differences in System 2 *ability* would predict performance. Experiment 3 was designed with this hypothesis in mind. Evans and Stanovich (2013) identified working memory as a defining System 2 ability, and an earlier discussion reviewed how tests such as the Cognitive Reflection Test and Raven Progressive Matrices are likely confounded theoretically as measures of cognitive ability. Working memory ability, by contrast, is a distinct measure of general cognitive processing ability, and is known to predict a wide range of complex outcomes (Conway, Kane, Bunting, Hambrick, Wilhelm, & Engle, 2005). Indeed, there are several precedents for looking at working memory as an underlying factor in reasoning and judgments under uncertainty (e.g., Capon, Handley, & Dennis, 2003; Copeland & Radvansky, 2004). Shi et al. (2006) and Yin et al. (2020) have in fact specifically found that working-memory ability predicts Bayesian reasoning performance, based on this NSA view interpretation. Experiment 3 uses a reading span test of working memory that is both performance-based and completely without numbers or mathematical calculations.

Methods

Participants

Data were collected from 239 undergraduates at a large public university, using the same recruitment process as in Experiment 1. The target sample size for this experiment was the same as for Experiment 2. After eliminating people

who did not complete the entire study (n = 39) and finished the entire study in less than 5 min (n = 0), the remaining 200 participants were used for analyses. The average age of participants was 19.32 years (SD = 1.88), and 67 of the participants (33.5%) were male.

Materials and procedures

The research used the same overall procedure as Experiments 1 and 2. The first section of the study consisted of a few demographic questions and four Bayesian reasoning tasks that systematically varied in terms of numerical format and presence of a pictorial aid. In order to comprehensively deal with any effects of context stories on Bayesian reasoning, Experiment 3 used a Latin-square design to counterbalance the numerical/tense formatting and the picture presence, relative to the context story used (see Brase, in press). An overall measure of Bayesian reasoning based on each participant's number of tasks correctly answered had a mean of 1.59 (*SD* = 1.20).

The second section consisted of four individual difference measures, presented in random order and with randomized item orders within each measure. One of these was a measure of numerical literacy (the MCQ-GNS; Hill et al., 2019) and another was a measure of visuospatial ability (the Paper Folding Test; Ekstrom et al. 1976). The MCQ-GNS showed acceptable reliability (Cronbach's α = .719) with a mean of 8.30 and standard deviation of 2.33, and the Paper Folding Test showed very good reliability (Cronbach's α = .877), a mean of 11.78, and a standard deviation of 4.94. The remaining two measures were the Set Theory Test (STT) developed in Experiment 1 and the Set Theory Test II (STT2) developed in Experiment 2. These both had very good reliability (Cronbach's α = .912 and .882, respectively) and the STT had a mean of 5.39 and standard deviation of 4.13, whereas the STT2 had a mean of 5.21 and standard deviation of 1.61.

The third section consisted of an entirely non-numerical measure of working memory; a reading span task adapted from Conway et al. (2005). This task consists of 15 trials, during each of which the participant is given individual letters to remember while performing a secondary task of reading and verifying whether sentences are semantically and syntactically correct. Each trial consists of between three to seven individual letters and interspersed sentence verification tasks, and the participants are asked to recall the individual letters in the order they were given. Scoring of this measure can be done in a few ways, but the most recommended method is a partial credit unit score (Conway et al., 2005), which is the mean proportion of letter elements recalled per trial. This measure overall had very high reliability (Cronbach's α = .941), a mean of 89.3% and a standard deviation of 16.9%.

Results

Performance across all the individual difference measures were significantly correlated (see Table in analyses in the Online Supplemental Material), so the subsequent analyses evaluated possible multicollinearity issues. An initial simultaneous entry multiple regression with all the potential predictor variables was statistically significant ($F(5,194) = 14.412$, $p < .001$), with an overall adjusted R^2 of .252. As in Experiments 1 and 2, the significant predictors were visuospatial ability (the Paper Folding Task, $p < .001$), and numeracy ($p = .001$). The variance inflation factors were all between 1.28 and 1.63, indicating that the predictors are moderately correlated but not problematically multicollinear (see analyses in the Online Supplemental Material for full details).

Following this, two separate hierarchical multiple regressions were run, as in the prior experiments. Both analyses entered working memory span first as a general processing ability measure. Working memory span, indeed, was a significant predictor of reasoning performance when entered first (adjusted R^2 of .043; $F(1,198) = 9.987$, $p = .002$).

At the second level either the visuospatial and numeracy measures were entered or the nested sets thinking measures were entered (Table 3). When the numeracy and visuospatial predictors were entered as the second step, the adjusted R^2 significantly increased to .255. At this step working memory span was no longer was a significant predictor, replaced by visuospatial ability and numeracy (both $p < .001$). The variance accounted for was not improved at the third step with the addition of nested sets ability predictors and the significant predictors remained only visuospatial ability ($p < .001$) and numeracy ($p = .001$).

When the nested sets predictors were entered as the second step (Table 3), the overall adjusted R^2 significantly increased to .101, driven by the SetTheory Scale ($p = .001$) and not working memory span any longer ($p = .113$). At the third step, however, the STT scale also fell out of statistical significance with the addition of numeracy and visuospatial predictors, and the variance accounted for a significant increase again (to .252). Both models, of course, end with all the same predictors, converging with the simultaneous entry model results at the third step ($F(5,194) = 14.412$, $p < .001$, adjusted R^2 of .252). The full regression results are provided in the analyses in the Online Supplemental Material.

Considering just participants who were half a standard deviation above the mean on both significant predictors (Paper Folding Task and General Numeracy Scale), those 51 people gave correct Bayesian responses 59.8% of the time. At the other extreme, participants half a standard deviation below the mean on both significant predictors ($n = 36$) gave correct Bayesian responses 16.7% of the time.

Table 3 Summary of results from hierarchical multiple regression analyses of Bayesian reasoning task performance, entering potential predictors at different levels per hypotheses

	R	R^2	Adjusted R^2
Numeracy and visuospatial ability first			
1: (Constant), Working Memory Span	.219	.048	**.043**
2: Model 1 + General Numeracy, Paper Folding	.516	.266	**.255**
3: Model 2 + Sets Reasoning Tests 1 and 2	.520	.271	**.252**
Nested sets ability first			
1: (Constant), Working Memory Span	.219	.048	**.043**
2: Model 1 + Sets Reasoning Tests 1 and 2	.338	.114	**.101**
3: Model 2 + General Numeracy, Paper Folding	.520	.271	**.252**

Discussion

About 30–35% of the variance in Bayesian reasoning performance can be accounted for by individual differences in numeracy and visuospatial ability. This is comparable with the finding of Brase and Hill (2017; just over 30% of the variance). Also consistent with prior findings, numerical literacy and spatial ability measures were correlated (along with other individual differences measures), but these do not appear to present multicollinearity problems in statistical analyses.

Theoretical implications

Some versions of the NSA view predict that a predisposition to think deliberatively (e.g., using System 2 processes) or have a particular thinking style (e.g., in terms of fleshing out models) is a key factor in terms of individual differences in Bayesian reasoning performance. There are, indeed, statistically significant relationships between reasoning performance and both need for cognition (Experiment 1) and the Rationality subscale of the Rational/Experiential Inventory (Experiment 2). These relationships, however, and are subsumed by numeracy and visuospatial ability.

Another version of the NSA view (e.g., Johnson-Laird et al., 1999) is that ability to do well in Bayesian reasoning is fundamentally related to abilities to mentally model conceptual set-theoretic situations, and thus individual differences in this ability should significantly account for variations in Bayesian reasoning. Several measures of ability to model and work with nested sets can indeed predict some additional variation in Bayesian reasoning performance. The predictive utility of nested sets modeling ability disappears, however, once numeracy and visuospatial ability are also included as predictors.

Yet another version of the NSA view (e.g., Sirota et al., 2014; Yin et al., 2020) is that ability to understand and reason about nested sets is a reflection of general cognitive processing ability. If one takes this view, then individual differences in general cognitive processing ability (separately measured) should dominate in accounting for variations in Bayesian reasoning performance. Cognitive processing ability (assessed with a working memory span that has no important numerical or visuospatial characteristics) can, in fact, predict some amount of Bayesian reasoning performance. Once again, though, that predictive ability is eliminated when numeracy and visuospatial ability are added as predictors. This result is inconsistent with the NSA view that rests on a general System 2 abilities explanation of individual differences in predicting Bayesian task performance.

Because "general ability" is a particularly broad concept, it is possible that other measures exist that could prove to be superior predictors of Bayesian reasoning (e.g., measures of intelligence, either in components or overall). The present dominance of numeracy and visuospatial ability, at least thus far, as predictors of the variation in Bayesian reasoning performance supports the ecological rationality view.

Applied implications

The repeated findings that numerical literacy and visuospatial skill are key underlying traits for successful Bayesian reasoning have implications for both selection and training of individuals who need to be able to engage in such activities. In terms of selection, these are key individual differences that can be used for evaluating who is better equipped to occupy positions that involve understanding and updating conditions based on incoming new information. For people already in such positions or aspiring to such positions, training in not only Bayesian reasoning specifically (e.g., Sedlmeier & Gigerenzer, 2001) but also the underlying skills of numeracy and visuospatial ability can improve their skills in this domain.

More broadly, these findings indicate pathways for improving educational statistics efforts overall. Building stronger foundations in numerical literacy and visualization of statistical situations can make later learning of more advanced

statistical inference topics – such as Bayesian reasoning – more successful.

Limitations and further issues

A converging result from here (Experiment 1) and Sirota et al. (2014, Experiment 1) is that the CRT was a predictor of Bayesian reasoning performance. The present research found this result with an expanded version of the CRT and without using a median split of CRT scores. The importance of this result is unclear and requires further research. As noted earlier, there continues to be debate about what exactly the CRT is a measure of. The CRT was used here (and in Brase & Hill, 2017) as a measure of numerical literacy, whereas Sirota et al. (2014) used it as a measure of "general reasoning." The correlation between the CRT and Bayesian reasoning drops substantially once numeracy is controlled for, but is still significant ($r_{partial}$= .271, p > .001), so there may be some other dimension(s) of the CRT that continue to be related to Bayesian reasoning performance. Further developments on the measurement of cognitive reflection ability (e.g., Sirota et al., 2018) may help to clarify both what the CRT is assessing and how it relates to other cognitive abilities such as Bayesian reasoning.

Based on Experiment 1, subsequent studies used specific measures of numeracy (the MCQ-GNS) and visuospatial ability (the Paper Folding task). An unexplored question is why these measures were the best predictors of the variation in Bayesian reasoning performance, or indeed if better predictors of numeracy and visuospatial individual differences exist.

Ability to mentally model conceptual set-theoretic situations, as measured with three new instruments developed here, was not related to Bayesian reasoning. One could possibly argue all of these measures of nested sets ability are somehow not validly measuring that ability, and further validation of these measures certainly would be useful, but an argument of this sort also begs the question of what would then be an acceptable assessment of set-theoretic modelling ability.

Finally, the current research helps to clarify a question that has been raised in the past (Brase et al., 1998): if ecologically rational minds are well designed for processing naturally sampled frequencies, why is Bayesian reasoning performance not higher? The answer seems to be that our minds were designed to operate in a rich world of actual events, objects, and locations; it is in some ways remarkable that they work as well as they do on paper and pencil problems. This understanding helps to address another possible question; why do individual differences in numeracy and visuospatial ability predict variation in Bayesian reasoning if the mind is purportedly so well adapted to reason with naturally sampled frequencies? These individual differences appear to be associated with the translational skills of taking the received text/pictorial information and making the mental connections with adaptations capable of calculating the relevant responses. Fortunately, these are skills that not only exist as individual differences, but they also are skills that can be developed with effort.

Acknowledgements This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The author has no conflicts of interests to declare. The experiments were not preregistered, and data and materials for all experiments are available at https://osf.io/cb5yh/?view_only=b253f3f251884eae89134270f3c2aee0.

References

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, *30*(3), 241–254. https://doi.org/10.1017/S0140525X07001653

Brase, G.L. (in press). What facilitates Bayesian reasoning? A crucial test of ecological rationality versus nested sets hypotheses. *Psychonomic Bulletin & Review*.

Brase, G.L., Cosmides, L. & Tooby, J. (1998). Individuation, counting, and statistical inference: The roles of frequency and whole object representations in judgment under uncertainty. *Journal of Experimental Psychology: General, 127*, 3-21. https://doi.org/10.1037/0096-3445.127.1.3

Brase, G.L. & Hill, W.T. (2017). Adding up to Good Bayesian Reasoning: Problem Format Manipulations and Individual Skill Differences. *Journal of Experimental Psychology: General, 146*, 577-591. https://doi.org/10.1037/xge0000280

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42,* 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Cacioppo, J. T., & Petty, R. E. (1984). The need for cognition: Relationships to attitudinal processes. In R. P. McGlynn, J. E. Maddux, C. Stoltenberg, & J. H. Harvey (Eds.), *Social perception in clinical and counseling psychology*. Lubbock, Texas Tech University Press.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48,* 306–307. https://doi.org/10.1207/s15327752jpa4803_13

Campitelli, G., & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory and Cognition, 42*, 434–447. https://doi.org/10.3758/s13421-013-0367-9

Capon, A., Handley, S., & Dennis, I. (2003). Working memory and reasoning: An individual differences perspective. *Thinking and Reasoning, 9*, 203-244. https://doi.org/10.1080/13546781343000222

Casscells, W., Schoenberger, A., & Grayboys, T. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine, 299,* 999–1000. https://doi.org/10.1056/NEJM197811022991808

Chapman, G. B., & Liu, J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgment and Decision Making*, *4*(1), 34–40.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring Risk Literacy: The Berlin Numeracy Test. *Judgment and Decision Making*, *7*, 25–47.

Colom, R., Escorial, S., & Rebollo, I. (2004). Sex differences on the Progressive Matrices are influenced by sex differences on spatial ability. *Personality and Individual Differences, 37,* 1289–1293.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks:

A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. https://doi.org/10.3758/BF03196772

Copeland, D. & Radvansky, G. (2004). Working memory and syllogistic reasoning. *Quarterly Journal of Experimental Psychology, 57*, 1437-1457. https://doi.org/10.1080/02724980343000846

DeShon, R. P., Chan, D., & Weissbein, D. A. (1995). Verbal overshadowing effects on Raven's Advanced Progressive Matrices: Evidence for multidimensional performance determinants. *Intelligence, 21*, 135–155.

Ekstrom, R.B., French, J.W., Harman, H.H. (1976). *Manual for kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

Evans, J. S., Handley, S. J., Perham, N., Over, D. E., & Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. *Cognition*, *77*(3), 197–213. https://doi.org/10.1016/S0010-0277(00)00098-6

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*, 223–241. https://doi.org/10.1177/1745691612460685

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.

Frederick, S. (2005). Cognitive reflection and decision-making. *Journal of Economic Perspectives, 19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Garcia-Retamero, R., Galesic, M., & Gigerenzer, G. (2010). Do icon arrays help reduce denominator neglect? *Medical Decision Making, 30*, 672-684. https://doi.org/10.1177/0272989X10369000

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. https://doi.org/10.1016/j.socscimed.2013.01.034

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*, 53–96. https://doi.org/10.1111/j.1539-6053.2008.00033.x

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

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*, 2343–68. https://doi.org/10.1080/17470218.2012.687004

Hill, W.T., Brase, G.L., Kenney, K. (2019). Developing a better and more user-friendly numeracy scale for patients. *HLRP: Health Literacy Research and Practice, 3*(3):e174-e180. https://doi.org/10.3928/24748307-20190624-01

Hoffrage, U., & Gigerenzer, G. (1998). Using Natural Frequencies to Improve Diagnostic Inferences. *Academic Medicine, 73*, 538-540.

Hoffrage, U., Hafenbrädl, S., & Bouquet, C. (2015). Natural Frequencies Facilitate Diagnostic Inferences of Managers. *Frontiers in Psychology, 6*: 642. https://doi.org/10.3389/fpsyg.2015.00642

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

Irwin, J. R. & McClelland, G. H. (2003). Negative effects of dichotomizing continuous predictor variables. *Journal of Marketing Research, 40*, 366–371. https://doi.org/10.1509/jmkr.40.3.366.19237

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

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N., Khemlani, S. S., & Goodwin, G. P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences*, *19*(4), 201–214. https://doi.org/10.1016/j.tics.2015.02.006

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

Kellen, V., Chan, S., & Fang, X. (2013). Improving user performance in conditional probability problems with computer-generated diagrams. In M. Kurosu (Ed.), *Human-Computer interaction: Users and contexts of use* (pp. 183-192). Berlin, Germany: Springer Berlin Heidelberg.

Lesage, E., Navarrete, G., & De Neys, W. (2013). Evolutionary modules and Bayesian facilitation: The role of general cognitive resources. *Thinking & Reasoning*, *19*(1), 27–53. https://doi.org/10.1080/13546783.2012.713177

Lindsey, S., Hertwig, R., & Gigerenzer, G. (2003). Communicating statistical DNA evidence. *Jurimetrics, 43*, 147163. http://www.jstor.org/stable/29762803.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*, 19–40. https://doi.org/10.1037//1082-989X.7.1.19

Macchi, L. (2000). Partitive formulation of information in probabilistic problems: Beyond heuristics and frequency format explanations. *Organizational Behavior and Human Decision Processes*, *82*(2), 217–236. https://doi.org/10.1006/obhd.2000.2895

McDowell, M. & Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychological Bulletin. 143*, 1273-1312. https://doi.org/10.1037/bul0000126.

Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology, 76*, 972-987. doi: https://doi.org/10.1037/0022-3514.76.6.972

Peters, M. & Battista, C. (2008). Applications of mental rotation figures of the Shepard and Metzler type and description of a mental rotation stimulus library. *Brain and Cognition*, *66*, 260-264. https://doi.org/10.1016/j.bandc.2007.09.003

Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R. & Richardson, C. (1995). A Redrawn Vandenberg & Kuse Mental Rotations Test: Different Versions and Factors that affect Performance. *Brain and Cognition, 28*, 39-58. https://doi.org/10.1006/brcg.1995.1032

Pighin, S., Gonzalez, M., Savadori, L., & Girotto, V. (2015). Improving public interpretation of probabilistic test results: Distributive evaluations. *Medical Decision Making, 35,* 12–15. https://doi.org/10.1177/0272989X14536268

Raven, J. C., Court, J. H., & Raven, J. (1977). *Manual for advanced progressive matrices (Sets I & IT).* London: H. K. Lewis & Co.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130*(3), 380–400.

Shi, Z., Qiu, J., & Zhang, Q. (2006). Facilitating Effect of Transparent Nested-Sets Relations on Bayesian Reasoning. *Acta Psychologica Sinica, 38*, 833-840. https://doi.org/10.1016/S0379-4172(06)60092-9

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. https://doi.org/10.3758/s13423-013-0464-6

Sirota, M., Kostovičová, L., Juanchich, M., Dewberry, C., & Marshall, A.C. (2018). Measuring Cognitive Reflection without Maths: Developing and Validating the Verbal Cognitive Reflection Test. Preprint available at https://osf.io/xehbv/.

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

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

Stanovich, K. E., & West, R. F. (1998a). Individual Differences in rational thought. *Journal of Experimental Psychology: General*, *127*(2), 161–188. https://doi.org/10.1037//0096-3445.127.2.161

Stanovich, K. E, & West, R. F. (1998b). Who uses base rates and P(D/-H)? An analysis of individual differences. *Memory and Cognition*, *26*(1), 161–179.

Stanovich, K E, & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–665; discussion 665-726.

Stanovich, K. E, & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*(4), 672–695. https://doi.org/10.1037/0022-3514.94.4.672

Thomson, K.S. & Oppenheimer, D.M. (2016). Investigating an alternate form of the cognitive reflection test, *Judgment and Decision Making, 11*, 99-113.

Toplak, M.E., West, R.F. & Stanovich, K.E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning, 20*, 147–168. https://doi.org/10.1080/13546783.2013.844729

Torrens, D., Thompson, V. A., & Cramer, K. M. (1999) Individual Differences and the Belief Bias Effect: Mental Models, Logical Necessity, and Abstract Reasoning. *Thinking & Reasoning, 5*, 1-28. https://doi.org/10.1080/135467899394066

Vigneau, F., & Bors, D. A. (2005). Items in context: assessing the dimensionality of Raven's Advanced Progressive Matrices. *Educational and Psychological Measurement, 65*, 109–123.

Yin, L., Shi, Z., Liao, Z., Tang, T., Xie, Y., & Peng, S. (2020). The Effects of Working Memory and Probability Format on Bayesian Reasoning. *Frontiers in Psychology*, 11:863. https://doi.org/10.3389/fpsyg.2020.00863