Structural mapping in statistical word problems: A relational reasoning approach to Bayesian inference

Journal:	*Psychonomic Bulletin & Review*
Manuscript ID	PBR-BR-15-340.R1
Manuscript Type:	Brief Report
Date Submitted by the Author:	n/a
Complete List of Authors:	Johnson, Eric; University of Barcelona, Psicologia Bàsica Tubau, Elisabet; University of Barcelona, Psicologia Bàsica
Keywords:	Bayesian inference, natural frequencies, numeracy, relational reasoning

Structural mapping in statistical word problems:

A relational reasoning approach to Bayesian inference

Eric D. Johnson [a,b,*] **& Elisabet Tubau** [a,b]

[a] Department of Basic Psychology, University of Barcelona, Barcelona, Spain

[b] IR3C, University of Barcelona, Barcelona, Spain

[*] Corresponding Author:

Eric D. Johnson

Departament de Psicologia Bàsica, Facultat de Psicologia

Universitat de Barcelona

Passeig de la Vall d'Hebron, 171

08035 Barcelona (Spain)

Phone: +34 93 312 51 40

Fax: +34 93 402 13 63

Email: eric.johnson@ub.edu

Acknowledgements

This research was supported by grants from the Generalitat de Catalunya (FI-DGR 2011) and the Spanish Ministry of Economics and Competitiveness (PSI2012-35703).

Abstract

Presenting natural frequencies facilitates Bayesian inferences compared to using percentages. Nevertheless, many people, including highly educated and skilled reasoners, still fail to provide Bayesian responses on these computationally simple problems. We show that the complexity of relational reasoning (e.g. structural mapping between presented and requested relations), can help explain remaining difficulties. With a non-Bayesian inference which required identical arithmetic but afforded more direct structural mapping, performance was universally high. Furthermore, reducing the relational demands of the task with questions which directed reasoners to use the presented statistics, compared with questions which prompted the representation of a second, similar sample, significantly improved reasoning. Distinct error patterns were also observed between these presented- and similar-sample scenarios, which suggested differences in relational reasoning strategies. On the other hand, while higher numeracy was associated with better Bayesian reasoning, higher numerate reasoners were not immune to the relational complexity of the task. Together, these findings validate the relational reasoning view of Bayesian problem solving, and highlight the importance of considering not only the presented task structure, but also the complexity of the structural alignment between presented and requested relations.

Keywords: Bayesian inference, natural frequencies, relational reasoning, numeracy, question form, structural mapping

1. Introduction

Educated adults are notoriously poor Bayesian reasoners with explicit numerical information (for recent review see Johnson & Tubau, 2015). While presenting statistical information as natural frequencies is the most widely agreed facilitator of Bayesian inferences (Gigerenzer & Hoffrage, 1995), many reasoners still fail to solve these problems. Why does Bayesian-like reasoning remain so difficult, even after providing reasoners natural frequencies? In this paper we address this question by viewing Bayesian reasoning as a case of relational reasoning, which requires the comparison of role-based structural relations across multiple mental representations (Halford, Wilson & Phillips, 1998, 2010; Holyoak, 2012).

The medical diagnosis problem represents a typical Bayesian reasoning task (Table 1). In this problem, information is presented regarding the base rate of having a disease and the likelihood of testing positive with a diagnostic test. As illustrated in Figure 1, this information can be represented structurally as a series of nested sets. From this information, the standard Bayesian question—(H|D)—asks reasoners to compute the expected number of infected people (the hypothesis, H) given a positive test result (the data, D). The solution, given natural frequencies, can be represented with the Bayesian equation:

$$(H|D) \ = \frac{(H\&D)}{(D)} = \frac{(H\&D)}{(H\&D) + (\neg H\&D)} = \frac{16}{16+24} = \text{"}16\ out\ of\ 40\text{"}$$

Problem Data:

A screening test is being studied to detect the presence of a new virus. The test is not perfect, however, as can be seen in the following data:

[100 people participated in the study:**]** 20 of them were infected with the virus, and 80 were not infected. Among those infected with the virus, 16 had a positive reaction to the test. Among the people not infected with the virus, 24 also had a positive reaction to the test.

Questions:

	p(D): ***Total Data***	***p***(H|D): ***Bayes***
Similar-Sample	Imagine the test is given to a new group of **(**100**)** people with similar characteristics to the study above. Among all of the people who participate in the study, how many of them would you expect to have a positive reaction to the test? __ out of __	**(**Imagine the test is given to a new group of 100 people with similar characteristics to the study above.**)** Among those who have a positive reaction to the test, how many would you expect to be infected? __ out of __
Presented-Sample	Among all of the people who participated in the study, how many of them had a positive reaction to the test? __ out of __	Among those people who had a positive reaction to the test, how many of them were actually infected? __ out of __
Solutions:	"(16+24) *out of* 100"	"16 *out of* (16+24)"

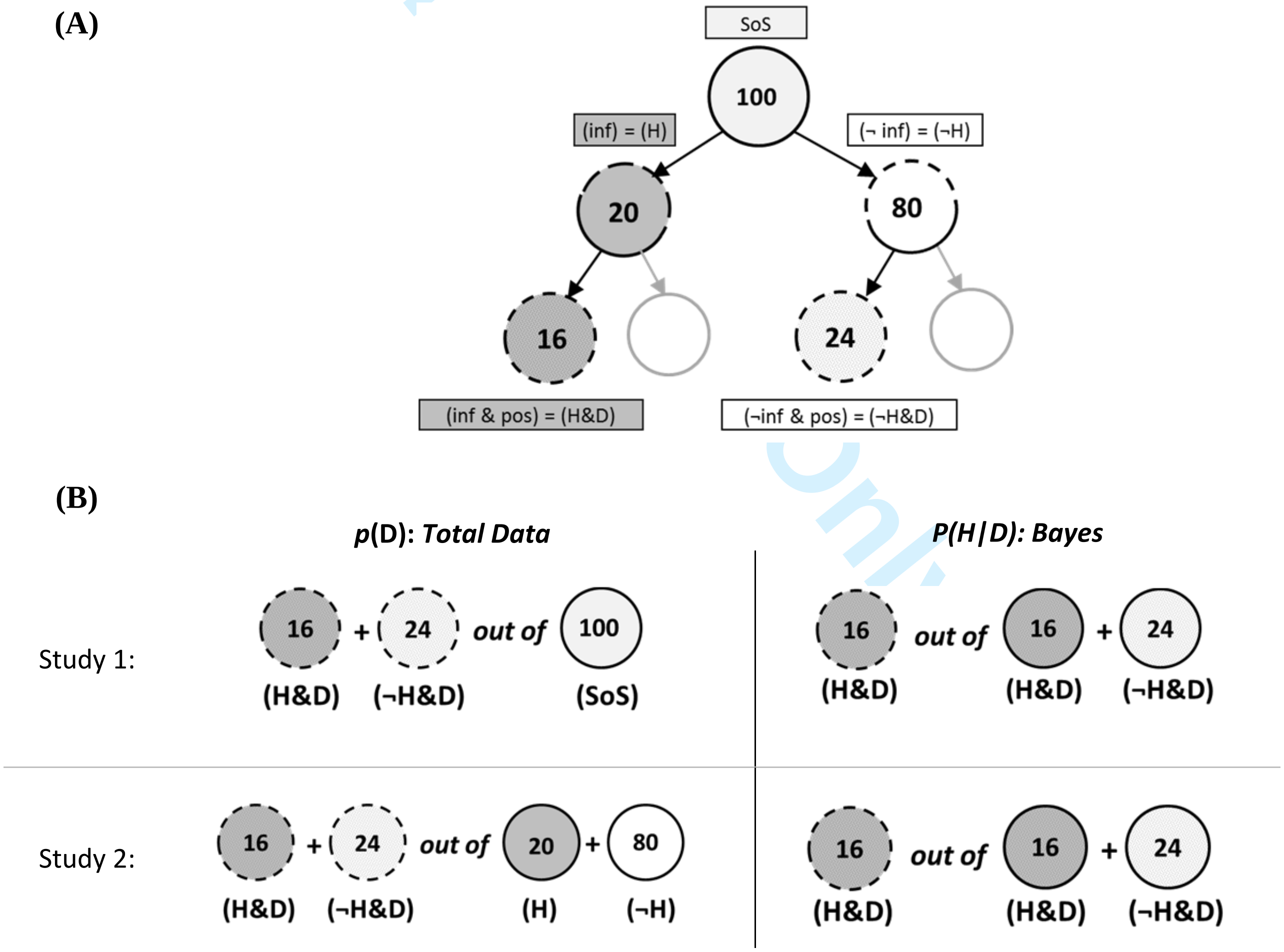
Table 1. Study 1 was a between-subject design, with each participant answering one of the four questions. Study 2 used a within-subject design, with each participant sequentially answering the p(D) and then p(H|D) questions in either the similar- or presented-sample framing. In study 2, the phrase in **[**brackets**]** was replaced with: "Among the people who participated in the study," and the information in **(**parentheses**)** was completely removed from the problem.

Intense theoretical debate persists regarding the mechanism with which natural frequencies facilitate performance over, for example, percentage formats (Barbey & Sloman, 2007; Brase & Hill, 2015). The most widely held view is some form of the nested sets theory.

While details vary among theorizers, this view is characterized by two primary claims: (1) making transparent the nested partition structure of a problem facilitates Bayesian reasoning, and (2) this facilitation arises out of a general reasoning mechanism operating over transparent set relations (e.g. Barbey & Sloman, 2007; Lesage, Navarrete, & De Neys, 2013; Sirota et al., 2014). While we generally agree with both claims, the nature of these set reasoning mechanisms has yet to be clearly described (see Mandel, 2007), and no paper has directly addressed the difficulties that remain after problem structures have been made "transparent" with natural frequencies. We argue that the main remaining difficulty is the misalignment of the relational roles between presented and requested data.

On our account, Bayesian reasoning can be understood as a special case of relational reasoning, which depends on understanding role-based structural relations (Gentner & Markman, 1997; Halford et al, 1998, 2010; Holyoak, 2012). Most reasoning errors arise from making superficial associations at the expense of understanding structural relations. For example, a common response to the above Bayesian question (p(H|D) in Table 1) is the hit rate (16 out of 20), which might arise from a direct association of common concepts in presented and requested relations (e.g. being infected, testing positive). Accurately carrying out the more complex Bayesian inference requires not only understanding the surface similarity of these elements, but also understanding the specific roles these categories play within a specified relation (e.g. reference class, focus subset; see Figure 1). Hence, an implication of the relational reasoning framework is that the fit of the structural alignment between the information presented in the text and requested in the question should predict reasoning performance (for extended discussion see Gentner & Markman, 1997; also Johnson & Tubau, 2015).

FIGURE 1. The information is Table 1 represented structurally as a series of nested set relations. (A) The structure of the presented information and relations to the Bayesian equation. (B) The requested relations needed to solve the p(D) and p(H|D) questions in studies 1 and 2. The relational roles played by the different elements (e.g. H, D) are illustrated with **solid lines (= reference class)** and **dotted lines (= focal subset)**. Note that H and ¬H play both roles in the presented data. In p(D) relational roles remain the same between presented and requested data, while they are misaligned with the p(H|D) question. H = infected; ¬H = not infected; H&D = infected and test positive; ¬H&D = not infected and test positive; SoS = superordinate set.



Alignment Hypothesis

From the perspective of relational reasoning, solving the Bayesian question would be difficult due to the change in the relational role of the subsets required to compute the reference class. That is, the new reference class—(D, '*all positive tests'*)—must be computed from previously presented focal subsets (H&D = '*positive among infected'* and ¬H&D = '*positive among not infected*'), from which the new focal subset (H&D) is selected (see Figure 1). Some studies have suggested that people do not understand that they must use both the focal (H&D) and alternative (¬H&D) hypotheses to compute the relevant reference class of the posterior ratio (D), where people tend to neglect the alternative hypothesis (see Evans et al, 2000; Girotto & Gonzalez, 2001; Krynski & Tenenbaum, 2007). Based on the relational reasoning approach, we hypothesized that the difficulty computing (D) is specific to its different role in the Bayesian relation (as reference class) relative to its role in the presented relations (as a pair focal subsets) rather than to the computation of (D) itself (*alignment hypothesis*).

To test this hypothesis, we created a non-Bayesian condition—*total data question*, p(D)—that required identical arithmetic as the Bayesian question, but where the elements (e.g. positive tests) played the same relational roles as in the presented data (Table 1). In this condition, the required reference class was the superordinate set (SoS) of the problem, while the new focal subset required the summing of two initial focal subsets (the frequencies corresponding to H&D + ¬H&D). Accordingly, the p(D) question also provides a control of participants' ability to select the particular subsets needed to compute the posterior reference class. The required computations are illustrated in the equation:

$$(D) \ = \frac{(D)}{(SoS)} = \frac{(H\&D) + (\neg H\&D)}{(SoS)} = \frac{16 + 24}{100} = \text{"}40\ out\ of\ 100\text{"}$$

Sample-Type Hypothesis

In addition to the mapping complexity between presented and requested relations, typical Bayesian tasks request participants to reason over a new, unspecified sample similar to the one presented (see Table 1, “similar-sample questions”). This requires the reasoner to infer the new statistics based on the presented statistics prior to carrying out the Bayesian inference. A series of informal observations in pilot studies in our lab suggested that simply asking participants to base their answer on the presented data greatly facilitated Bayesian responses. From the relational reasoning account, a question prompting a reasoner to imagine a similar sample would increase processing demand due to the need to maintain different samples (the presented sample, and the imagined similar sample) in order to infer the corresponding statistics. Likewise, instructing participants to base their answer on the presented data would facilitate exact Bayesian responses by eliminating the need to construct a second representation and perform the corresponding mapping. We test this *sample-type hypothesis* by directly instructing participants to reason over the sample of individuals presented in the text (presented-sample question) or over a new, similar sample of individuals (similar-sample question) (Table 1).

Alternatively, asking to imagine a similar sample might lead participants to base responses on an approximation or a compute-then-adjust strategy to accommodate the “similar” sample. In this case, more approximate estimations should be observed in the responses to both Bayesian and non-Bayesian questions. However, if, as suggested above, imagining a similar sample hampered performance due to increased relational reasoning demand, this effect should be stronger for the more complex Bayesian question. To distinguish between these possibilities, error analyses were used to gain insight into the processing strategies that individuals were engaging. Given that these hypothesized effects might also be moderated by individual skills, we measured participants' level of numeracy, commonly observed to influence probabilistic reasoning performance (e.g. Johnson & Tubau, 2013; McNair & Feeney, 2015).

2. Study 1

In study 1 we used a between-participant design to test the *sample-type hypothesis*—that Bayesian performance can be improved by instructing reasoners to use the data presented in the problem versus imaging a new similar sample—and the *alignment hypothesis*—that even with identical numerical computations requested from identically presented data, a specific difficulty will be observed with the Bayesian logic.

2.1. Method & Material

Based on an a priori power calculation to detect a medium sized effect (w=0.4), we aimed to include at least 60 participants for each condition (a conservative estimate considering previous research; e.g. McNair & Feeney, 2015; Sirota et al, 2014, 2015). Participants were 319 undergraduates from the University of Barcelona who had yet to receive instruction in Bayesian reasoning. Informed consent was collected and students received course credit for their participation. All participants received the virus test scenario and either the p(D) or p(H|D) question referring to either an imagined *similar* sample or the *presented* sample shown in Table 1. Following this, participants completed the 11-item Lipkus et al (2001) numeracy scale. Calculators were not allowed.

2.2. Results

Nine participants were removed from the study for indicating they had previously seen this type of problem or for not completing the task as requested. Results are therefore reported for the remaining 310 participants (mean age = 21.4, SD = 1.6). The mean numeracy score was 8.65 (median = 9, range = 3-11, SD = 1.86).

Reasoning Accuracy. Global results are summarized in Figure 2. Only exact Bayesian answers were counted as correct. A logistic regression was performed using the dichotomous coding of response (correct, incorrect) as the dependent variable, with the question (p(H|D), p(D)), sample type (similar, presented), and continuous numeracy score entered as predictors.

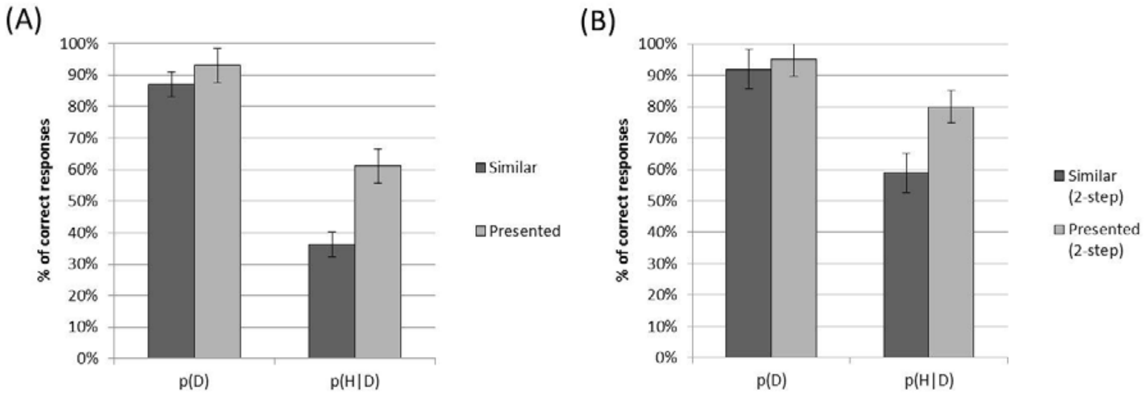


Figure 2. Proportion of correct responses with similar vs. presented sample questions, with the alignment manipulation, p(D) vs. p(H|D), as (A) between-subject in study 1, and (B) within-subject as a two-step manipulation in study 2. Error bars are standard errors.

Results revealed a significant main effect of sample type, $\chi^2(1) = 12.36$, $p < .001$, $e^{\beta} = 2.90$, 95% C.I. = 1.60 – 5.24, showing facilitated performance when the question directed reasoners to use the specific data presented in the problem. There was also a significant effect of question, $\chi^2(1) = 50.92$, $p < .001$, $e^{\beta} = 11.22$, 95% C.I. = 5.77 – 21.79, with ceiling performance observed with the p(D) question. Finally, as illustrated in Figure 3A, a significant effect of numeracy was also observed, $\chi^2(1) = 14.65$, $p < .001$, $e^{\beta} = 1.37$, 95% C.I. = 1.17 – 1.61, with higher numerate individuals better with both presented and similar scenarios. There were no significant interactions (all *p*s >.20).

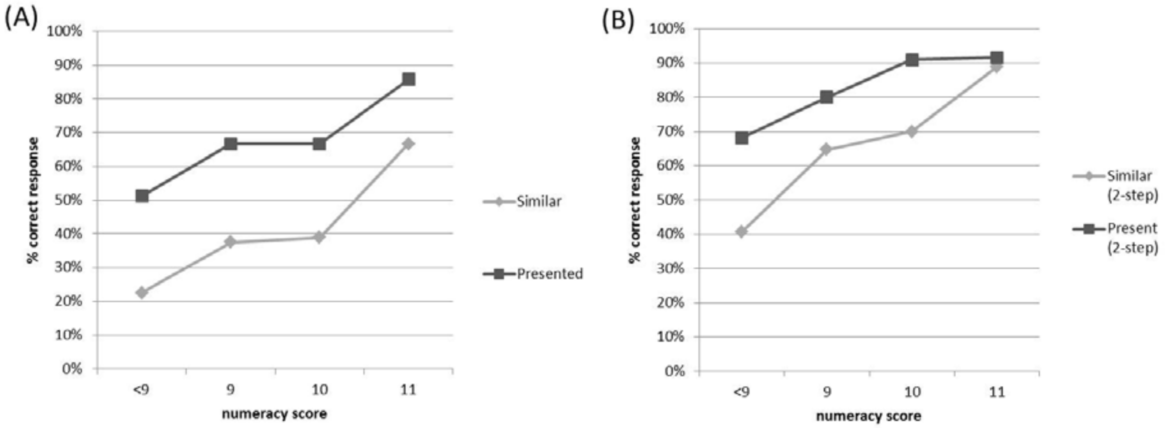


Figure 3. Performance with the Bayesian p(H|D) question according to numeracy score (A) in study 1, and (B) in study 2 with a two-step manipulation.

Error Analysis. Table 2 reports observed errors on the p(H|D) question. Errors in the *similar-sample* condition were widely distributed, however, over half of errors contained the superordinate value '100' as the reference class. In the *presented-sample* condition, less than a third of all errors contained '100'. This difference in responses containing '100' was significant, χ^2(N=153) = 11.47, p < .001, φ = .27. Furthermore, in the similar-sample scenario only a few hit-rate-only responses were observed, whereas this was the most frequent error in the presented-sample scenario, χ^2(N=153) = 5.91, p = .015, φ = .20 (see Table 2).

St.	*Sample Type*	(n)	Bayes (16/40)	Pre-Bayes (20/40)	HR (16/20)	BR (20/100)	Joint (16/100)	100-othr (xx/100)	Other
1	Similar	(69)	**36**	10	3	12	6	17	16
	Present	(84)	**61**	4	14	2	2	7	10
2	Similar (2-stp)	(63)	**59**	8	8	3	6	5	11
	Present (2-stp)	(60)	**80**	2	12	0	2	0	4

Table 2. Percentages of responses observed with the posterior p(H|D) question in studies (St.) 1 and 2. *Bayes* = exact Bayesian solution (H&D *out of* D); *Pre-Bayes* = correct denominator but selecting all of infected as numerator (H *out of* D). *HR* = hit rate (H&D *out of* H); *BR* = base rate (H *out of* 100); *Joint* = joint occurrence (H&D *out of* 100); *100-othr* = other numerators paired with superordinate set as denominator (xx *out of* 100).

Discussion. Results of this study supported the two primary hypotheses. First, we obtained an alignment effect: significantly better performance when the relational roles in the presented and requested set-subset data were aligned. Second, we observed a sample-type effect: improved performance by instructing participants to reason over the data presented in the problem (presented-sample question) compared with a similar sample of individuals (similar-sample question). The ceiling performance with the p(D) question, requiring the same computations as the more complex p(H|D) question, indicates that difficulties determining the Bayesian reference class (the total set of positive tests) are specific to its particular role in the posterior ratio. Errors analyses suggested that the lower performance in the similar sample scenario was not simply caused by more approximate reasoning strategies, since this would have also reduced exact responses in the p(D) question. Nevertheless, a large number of responses in the similar-sample scenario contained the value '100'. The presence of this value in the presented data and in the similar-sample scenario questions (see Table 1) could have both hampered the difficult question and enhanced the easier one. Study 2 attempted to solve this issue.

3. Study 2

In this study we provide a more stringent test of the sample-type and alignment hypotheses. As commented above, the inclusion of the '100' both in the data and in the question of the similar-sample scenario may have both interfered with the selection of relevant information for the p(H|D) question, and also facilitated performance for p(D), which together would undermine the observed alignment effect, in addition to weakening the sample-type effect. Therefore, in this study we removed all references to the superordinate set value '100' from the problem text and question. This both removed irrelevant information from the p(H|D) problem, and added an additional arithmetic step to the p(D) question (Figure 2). We also adopted a two-step design where participants were sequentially presented the p(D) and p(H|D) questions as a within-subject manipulation.

We again expected near-ceiling performance with the initial p(D) question. We hypothesized that these manipulations would facilitate Bayesian responding in both similar- and presented-sample conditions. A significant sample-type effect under these highly facilitatory conditions would be clear evidence that reasoning over similar samples requires additional processing compared to reasoning over presented samples.

3.1. Method & Materials

A new sample of 123 undergraduate students (mean age = 21.2, SD = 1.4) completed the p(D) question followed by p(H|D), both with either presented-sample or similar-sample framing, and also the numeracy scale. The alignment manipulation was therefore within-participant, while the sample-type manipulation was between-participant. All explicit mentions of '100' were removed from the problem as indicated in Table 1.

3.2. Results

Reasoning Accuracy. Global results are shown in Figure 2B. The mean numeracy score was 8.76 (median = 9, range = 3-11, SD = 1.71). With the p(D) question, ceiling performance was again observed with both sample-type scenarios, and a logistic regression run on this first step showed no differences between sample type, numeracy scores, and no interaction, all χ^2 (1) < 1.5, *ps* > .25. In contrast, a logistic regression on the p(H|D) response revealed significant main effects of both sample type, χ^2 (1) = 6.91, *p* = .009, e^{β} = .31, 95% C.I. = .13 – .74, and numeracy,

χ^2 (1) = 11.27, p= .001, e^{β} = 1.54, 95% C.I. = 1.20 – 1.98 (Figure 3B). The interaction was not significant, χ^2 (1) < 1, p > .75.

A within-subject McNemar's test also confirmed an alignment effect with both the presented-sample, McNemar χ^2(N=60) = 9.0, p = .004, φ = .46, and similar-sample, McNemar χ^2(N=63) = 17.64 9, p < .001, φ = .11, scenarios. To further explore this effect according to numeracy level, a median split was used to separate participants into higher and lower numeracy groups. For lower numerate individuals, a significant alignment effect was observed for both the similar-sample, McNemar χ^2(N=27) = 13.00, p < .001, φ= .29, and presented-sample, McNemar χ^2(N=22) = 6.00, p = .031, φ = .32, scenarios. For higher numerate reasoners, the alignment effect just reached significance with the similar-sample scenario, McNemar χ^2(N=36) = 5.33, p = .039, φ= .15, however, it disappeared in the presented-sample condition, McNemar χ^2(N=38) = 3.00, p = .25.

Error Analysis. Globally, the most common errors paralleled those found in the previous study (Table 2). In particular, although errors with the similar-sample condition were widely distributed, the use of '100' as the reference class was still more frequent than with the presented-sample scenario, χ^2(N=123) = 6.55, p = .010, φ = .23. Also as before, most of the errors in the presented-sample scenario corresponded to the hit rate (58% of errors, vs. 14% of errors with the similar-sample condition; χ^2(N=38) = 5.81, p = .016, φ= .39).

Discussion. The present study provides additional support for the *sample-type effect*. Even with two-step questions which guide reasoners through the necessary computations, instructing reasoners to use the presented data still facilitated Bayesian responses compared with similar samples. As discussed below, error analysis also suggested that the presented-sample enhanced the mapping between presented and requested relations. Nevertheless, an *alignment effect* was still observed for individuals lower in numeracy. Confirming previous findings, almost all participants in both conditions could use the presented information to accurately compute the total number of positive tests using the hit rate and false-positive rate, but many of them still failed to use this computation as the reference class for the Bayesian ratio. Together, this indicates a specific difficulty with the structural mapping required to supply the Bayesian response.

4. General Discussion

In these studies we asked why Bayesian-like reasoning remains so difficult even after clarifying the nested set structure of the problems with natural frequencies. On our account, statistical word problems require a type of relational reasoning, and therefore performance should be influenced by the relational complexity of the task. Solving the Bayesian p(H|D) question requires realizing that the relational roles of specific subsets presented in the text are changed in the question. With the non-Bayesian p(D) question, on the other hand, the relational roles of focal and references class are maintained between the presented and the requested information, providing a more direct structural mapping. Accordingly, compared to the p(D) question, the p(H|D) inference requires an added level of abstraction to notice that the *relational role is not fixed* and can vary with the form of the question.

The observed *sample-type effect* is also consistent with the relational reasoning account, providing clear evidence that there is a cost when people are instructed to reason over a similar sample compared to when instructions direct reasoners to use the presented data. Of interest, however, the similar-sample framing did not significantly impair reasoning with the p(D) question in study 2, and this was the case even with the added arithmetic step in this condition. This suggests that the misalignment between presented and requested relations is the primary relational burden for the Bayesian inference, which is made more complex with the added processing demands of the similar-sample question.

Error analyses revealed that the poorer performance in the similar-sample condition did not stem from a shift in strategy, but rather from a confusion in the text comprehension process and/or processing interference during the mapping from presented to similar sample. Analysis of the 'other' responses argued against the possibility that reasoners were using estimation strategies or computing the Bayesian response and then adjusting slightly for possible uncertainties in the newly imagined sample (*compute-then-adjust* strategy). The fact that virtually all participants answered p(D) with exactly "40 out of 100" also demonstrates that the similar-sample questions do not inherently invoke estimation or compute-then-adjust strategies. It might also be suggested that the prevalence of the superordinate value '100' indicates that participants are attempting to normalize responses. A careful review of protocols, however, argued against this possibility, with most of these responses showing drawn arrows or circles around the "100" in the text, or a simple summing of 80+20, rather than normalization procedures. Furthermore, the '100' in the denominator was roughly equally paired with numbers presented in or directly derived from the text (16, 20, 40), making the normalization explanation or the estimation strategy less likely. While these patterns indicate that the error variability did not result from the application of estimation strategies or normalization procedures, future studies could look to more carefully specify the type of processing burden driving this effect.

As predicted, there were fewer overall errors with the presented-sample scenarios, most of which were the hit-rate, or inverse fallacy (Villejoubert & Mandel, 2002). The hit rate might be explained by superficial strategies such as label matching (Evans, 1998) or, more specifically, by confusing the structural roles (reference class and focal subset) during the mapping between presented and requested relations (Holyoak & Koh, 1987). That is, participants might make the less effortful direct mapping between elements in the question and presented text (infected-positive) rather than the more demanding mapping requiring consideration of the relevant relational roles (total positive→infected). Hence, the more straightforward relational mapping promoted by the presented-sample scenario can explain both the increment of correct responses and the most frequent error found in this scenario.

Much work has gone into understanding why natural frequencies facilitate Bayesian performance. We were motivated by the general claims and empirical support for the nested sets theory to further explore why natural frequencies still remain so difficult for so many reasoners. Previous work has looked at both the phrasing of the presented information (e.g. Krynski and Tenenbaum, 2007; McNair& Feeney, 2015) and the form of the question (e.g. Girotto & Gonzalez, 2001; Pighin et al, 2015), but no study has specifically looked at the role-based structural compatibility explored in the present study. One recent proposal by Ayal and Beyth-Marom (2014) looked at the “compatibility” of presented and requested relations, however, their compatibility manipulations looked specifically at numerical aspects (formats, sample sizes), not structural compatibility. They also reported an effect of “mental steps” referring to the number of explicit calculations needed to solve the problem. In our work, p(D) and p(H|D) require the same number of numerical steps, and therefore our alignment effect demonstrates an additional burden not tied to explicit numerical transformations, namely, the relational reasoning required to map the misaligned structures. We believe the natural frequency format enhances relational reasoning (see also DeWolf, Bassok, Holyoak, 2015). However, we also make the stronger claim that problems using percentages could also be viewed as relational reasoning tasks requiring the alignment of role-based structured relations, in addition to the corresponding numerical transformations. Accordingly, the relational reasoning framework predicts that similar alignment effects should be observable in Bayesian problems using percentage formats. Future studies could test this hypothesis.

To conclude, while natural frequencies have been hailed as a facilitator of Bayesian inferences for twenty years now (Gigerenzer & Hoffrage, 1995), the fact that performance on these problems still remains so low has been little discussed. The present studies offer an explanation for this difficulty by viewing Bayesian word problems as a case of relational reasoning, which requires the comparison of structural relationships across different level of abstraction.

References

Ayal, S. & Beyth-Marom, R. (2014). The effects of mental steps and compatibility on Bayesian reasoning. *Judgment and Decision Making*, 9(3), 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.

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*.

DeWolf, M. Bassok, M., Holyoak, KJ (2015). Conceptual Structure and the Procedural Affordances of Rational Numbers: Relational Reasoning With Fractions and Decimals. *J of Exp Psych: Gen*, 144(1), 127–150.

Evans, J. St. B. T. (1998). Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking & Reasoning*, 4, 45-82.

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.

Gentner & Markman, A.B. (1997). Structure Mapping in Analogy and Similarity. *American Psychologist*, 52(1), 45-56.

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

Girotto, V. & Gonzalez, M. (2001) Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition*, 78, 247–76.

Halford, G.S., Wilson, W.H. & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, 21, 803–865.

Halford, G.S., Wilson, W.H. & Phillips, S. (2010). Relational knowledge: the foundation of higher cognition. *Trends in Cognitive Sciences*, 14(11), 497-505.

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 234-259). New York: Oxford University Press.

Holyoak, K.J. & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition*, 15(4), 332-340.

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(938). doi: 10.3389/fpsyg.2015.00938

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *J. Exp. Psych. Gen*eral. 136(3), 430–50. doi: 10.1037/0096-3445.136.3.430.

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

Mandel, D. R. (2007). Nested sets theory, full stop: Explaining performance on Bayesian inference tasks without dual-systems assumptions. *Behavioral and Brain Sciences*, *30*(03), 275-276.

McNair, S.J. & Feeney, A. (2015). Whose statistical reasoning is facilitated by a causal structure intervention? *Psychonomic Bulletin & Review*, 22(1), 258-64. doi: 10.3758/s13423-014-0645-y.

Pighin, S., Gonzalez, M., Savadori, L., and Girotto, V. (2015). Improving public interpretation of probabilistic test results: Distributive evaluations. *Medical Decision Making*, 35, 12-5. doi: 0.1177/0272989X14536268

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.

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

Villejoubert, G. & Mandel, D.R. (2002). The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. *Memory & Cognition*, 30(2), 171-178.