# Nested Sets and Natural Frequencies

Stephen H. Dewitt[1], Anne Hsu[2], David Lagnado[1], Saoirse Connor Desai[3], Norman E. Fenton[2].

[1]Department of Experimental Psychology, University College London, 26 Bedford Way, WC1H 0AP
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Rd, London E1 4NS
[3] Psychology Department, City University, London, EC1R 0JD

**Abstract**

Is the nested sets approach to improving accuracy on Bayesian word problems simply a way of prompting a natural frequencies solution, as its critics claim? Conversely, is it in fact, as its advocates claim, a more fundamental explanation of why the natural frequency approach itself works? Following recent calls, we use a process-focused approach to contribute to answering these long-debated questions. We also argue for a third, pragmatic way of looking at these two approaches and argue that they reveal different truths about human Bayesian reasoning. Using a think aloud methodology we show that while the nested sets approach does appear in part to work via the mechanisms theorised by advocates (by encouraging a nested sets representation), it also encourages conversion of the problem to frequencies, as its critics claim. The ramifications of these findings, as well as ways to further enhance the nested sets approach and train individuals to deal with standard probability problems are discussed.

**Keywords:** Nested Sets; Natural frequencies; Bayesian; Base rate neglect

A recent meta-analysis (McDowell & Jacobs, 2017) conclusively demonstrated that when a Bayesian word problem is presented according to natural frequency (NF) principles, normative responding increases relative to the 'standard probability' format (SP), with an average accuracy of around 24%. Both versions of the classic medical diagnosis problem can be seen below (statistical notation added).

**Standard probability format (individual chance):** The chance of breast cancer is 1% [P(Ca)] for women at age forty who participate in routine screening. If a woman has breast cancer, the chance is 80% [P(Po|Ca)] that she will get a positive mammography. If a woman does not have breast cancer, the chance is 9.6% [P(Po|¬Ca)] that she will also get a positive mammography. A woman in this age group had a positive mammography in routine screening. What is the chance that she actually has breast cancer [P(Ca|Po)]? ____%

**Natural frequencies:** 10 [F(Ca)] out of 1000 women at age forty who participate in routine screening have breast cancer. Out of the 10 women with breast cancer, 8 [F(Po&Ca)] will get a positive mammography. 95 [F(Po&¬Ca)] out of every 990 women without breast cancer will also get a positive mammography. Here is a new representative sample of women at age forty who got a positive mammography in routine screening. What proportion of these women do you expect to actually have breast cancer [P(Ca|Po)]? ___ %

We can see several differences between these formats. Most obviously, the NF format uses frequencies (indicated by the 'F' notation) rather than percentages / probabilities (P), but more importantly, the figures are not normalized. In the SP format, the figures are normalized by the use of a standard denominator (percentages are one way of achieving this with a hidden denominator of 100, but normalized frequencies with other denominators are also possible). This difference in normalization firstly has a known effect on the number of computations required to solve each problem. In an NF format there are thought to be only two computational steps[1]: (1) summing the number of individuals with a positive result and cancer F(Po&Ca) with the number of individuals with a positive result but no cancer F(Po&¬Ca) and then (2) dividing F(Po&Ca) by this sum. The same formula can be used if those same numbers are given in percentage or probability format.

$$\frac{F(Po\&Ca)}{F(Po\&Ca) + F(Po\&\neg Ca)} \; or \; \frac{P(Po\&Ca)}{P(Po\&Ca) + P(Po\&\neg Ca)}$$

However, normalized formats require an additional pre-step (you won't see any of these figures in the standard probability format to the left). P(Po&Ca) must itself first be calculated by multiplying the proportion of individuals with cancer who get a positive result (P[Po|Ca]) with the total proportion of individuals with cancer P(Ca). Similarly, P(Po&¬Ca) must be calculated by multiplying P(Po|¬Ca) with P(¬Ca). For example, to calculate the proportion of women without breast cancer and a positive result, we multiply the percentage of women without breast cancer (99%) by the percentage of those women who get a positive result (9.6%). This may be a trivial calculation for most, but crucially, the solver first has to have an accurate representation of the problem in order to know that we (A) need to calculate this figure to solve the problem and (B) should multiply these two particular values rather than using some other figures or operation to compute it.

As has been noted, in the natural frequency format, this figure is provided for us, which has widely been accepted as a potential confound by subverting the need for (A) entirely (however see Brase & Hill, 2015 for work suggesting this may not be an important factor). However, NF

[1] In fact, in some natural frequency versions, the final question is: 'How many of these women do you expect to actually have breast cancer? ___ out of ___' This reduces the computational steps further, to one only: calculating the total positives.

proponents (e.g. Gigerenzer and Hoffrage, 1995) tell the story the other way around: normalization is an artificial (and relatively recent) human construct which transforms problems from a natural and solvable format to an unnatural and difficult one. These authors propose that normalization adds an additional difficulty by changing the structure of the information from that which would be obtained through 'natural sampling' i.e. if we observed 1000 women one by one, taking note in each case whether they had cancer and whether they got a positive result. This information structure of the natural frequency format is thought to replicate the natural format that human beings experience in the world, and thus are predisposed in some way to work with, which is the true reason for the increased normative responding (Gigerenzer & Hoffrage, 1995).

One concrete change however is that when information is presented in this way, the denominator of F(Po&¬Ca) (990) matches F(¬Ca). Other authors (e.g. Evans, Handley, Perham, Over & Thompson, 2000; Sloman, Over, Slovak & Stibel, 2003) have therefore claimed that rather than this having anything to do with 'natural' formats, this simply makes the 'nested sets' structure of the problem transparent (e.g. that women with a positive mammography but no breast cancer are a subset of the larger group of women without breast cancer). Nested sets advocates argue that this set structure revelation should be considered the more ultimate cause. They have sought to demonstrate that any method which reveals the nested sets structure of the problem will be equally successful. One example, using normalized percentages for the false positive and negative rates like the SP format but framing these in terms of proportions of groups (PP) rather than individual chance (an approach developed by Macchi [2000]), can be seen below:

**Nested Sets (Proportion Percentages):** 10 F[Ca] out of 1000 women at age forty who participate in routine screening have breast cancer. Out of the women with breast cancer, 80% [P(Po|¬Ca)] will get a positive mammography. Out of those women without breast cancer, 9.6% [P(Po|¬Ca)] will also get a positive mammography. Here is a new representative sample of women at age forty who got a positive mammography in routine screening. What proportion of these women do you expect to actually have breast cancer [P(Ca|Po)]? ___ %

Macchi (2000) found an improvement in accuracy compared to an SP format, and no significant difference to an NF format. Following this and similar papers, NF proponents (Hoffrage, Gigerenzer, Krauss & Martignon, 2002) have argued that nested sets formats simply encourage solvers to construct an NF version of the problem for themselves, which is the ultimate reason for increased accuracy. This criticism seems all the more plausible for Macchi's format, given that unlike the standard probability format, it presented the base rate as a frequency. It is important to note however that Gigerenzer and Hoffrage (1995) originally theorized based on evolutionary grounds that the phenomena of neglecting base rates (P[Ca] and P[¬Ca]) during solution should generalize to non-NF formats because that information is not required for solution in an NF format, which people are adapted to:

*"Base rate information need not be attended to in frequency formats (Result 3).* If our evolutionary argument that cognitive algorithms were designed for frequency information acquired through natural sampling is valid, then base rate neglect may come naturally when generalizing to other information representations, such as the standard probability format (Gigerenzer & Hoffrage, 1995, pp. 29)

While the authors refer specifically to the standard probability format here, the key point is that in evolutionary history humans have never had to complete the 'pre-step' required in the normalized format, because information has always been presented to them in the natural frequency format (and in which they can compute the normative answer without using the base rates), and so they may lack the capacity to do this, regardless of whether that normalized format is presented in the SP way, or in Macchi's PP way. The simple fact that nested sets results defy this has been widely overlooked in the field, and in fact suggests a potential harmony between the two approaches, rather than a discord, at least at the pragmatic level. While people do indeed seem more capable of solving a Bayesian word problem in a natural frequency format, than in a standard normalized format, nested sets results show us that, with the right framing, people can solve normalized Bayesian problems too.

A preliminary aim of this paper is to replicate Macchi's approach, as it has only been demonstrated in a single experiment. Furthermore, it needs replication in a wider range of more ecologically valid situations, including with the base rate presented as a percentage (as mentioned, Macchi's original format used a frequency base rate unlike the SP format) and with non-whole numbers. These factors may be present in real-world contexts and may add sufficient complexity to undermine the value of the format. We also aim to test the format in both simple (all women with breast cancer get a positive result) and hard (some women with breast cancer get a false negative) problems as both versions have been used widely in the literature.

A more ambitious aim of this paper is to assist in settling the highly debated connection between nested sets and natural frequency formats. Over the past few years repeated calls have been made to resolve these differences between the two camps (Brase & Hill, 2015; McNair, 2015; Johnson & Tubau, 2015; McDowell & Jacobs, 2017). Given that these are fundamental questions about cognitive process, the same authors have repeatedly called for more process-focused experiments. While two previous experiments (Gigerenzer & Hoffrage 1995; Macchi, 2000) used a 'think aloud' (TA) approach (where participants record their thought processes while solving the problem) in both cases

this was only used to report the types of errors participants make. We aim to make greater use of this data to shed light on the following questions. Does the nested sets approach work, as claimed by its advocates, by encouraging a representation of e.g. P(Po&¬Ca) as a subset of P(¬Ca) at the first, de-normalization step? Does the nested sets approach encourage individuals to construct a natural frequency representation for themselves, as claimed by Hoffrage et al. (2002)? Which of these are predictive of success on the problem? Finally, what else can we learn about the mechanisms by which Macchi's nested sets approach achieves greater accuracy?

## Method

521 participants were recruited through Amazon MTurk (55.3% female; mean age = 34.2 [SD = 11.6]). The experiment had eight between-subjects conditions, using a 2 (standard probability [SP] vs proportion percentages [PP]) x 2 (simple vs hard) x 2 (whole vs decimal) design. The PP-hard-decimal condition can be seen below (with statistical notation, not shown to participants), and further materials and experimental data are available at https://osf.io/nd46g/. This is considered a decimal version because the product of computational step 1 (e.g. 10% x 76% = 7.6%) is a non-whole number.

*Every year the government advises women to take part in routine mammography screening using an X-ray machine to determine if they have breast cancer. Among women at age forty who participte in this routine screening 10% [P(Ca)] have breast cancer, while 90% [P(¬Ca)] do not. However, the screening test is not always accurate. Specifically, out of those women who have breast cancer, only 76% [P(Po|Ca)] will actually get a positive mammography. Furthermore, out of all of those women who do not have breast cancer, 15% [P(Po|¬Ca)] will also get a positive mammography. What percentage of women at age forty who get a positive mammography [P(Po)] in routine screening actually have breast cancer[P(Ca|Po)]? ___%*

Participants were also required to record their thought process in an open text box. They could only submit their numerical response after they had submitted their thought process. All qualitative analysis of the TA data was undertaken blind to condition. Analysis was coded by two authors separately, with over 90% agreement. Discrepancies were resolved through the decision of a third coder.

Participants were given a 'normative' label if their numerical response was within 1% of the Bayesian normative value. Beyond this however, we found seven participants who clearly demonstrated accurate reasoning, including all necessary computational steps, but made an arithmetic error. These participants were also labelled as normative. One of these participants was in the nested sets conditions, while six were in the standard probability conditions.

## Results

### General Results

The overall proportion of the sample providing the normative response for the experiment was 13.5% with an average of 9.0% for the SP conditions and 18.1% for the PP conditions. In Figure 1, normative proportions for all eight conditions can be seen.

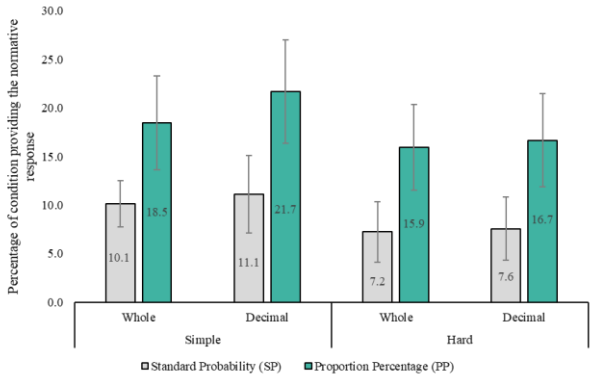


Figure 1. The percentage of participants providing the normative Bayesian answers across all eight conditions. Error bars represent one standard error.

A binary logistic regression (BLR) using 'normative response' as the dependent variable and the three condition-comparisons (SP vs PP; whole vs decimal; simple vs hard) as independent variables found a main effect for the SP-PP comparison (Wald $X^2 = 8.984$, p=.003), no main effect for the whole-decimal comparison (Wald $X^2 = .184$, p=.668) and no main effect for the simple-hard comparison (Wald $X^2 = 1.350$, p=.245). All subsequent analyses on 'condition' therefore compare SP to PP only.

### Nested Sets Representation

Across all conditions, 87 (16.7%) individuals expressed a 'nested sets representation' of the problem. For this classification, participants had to explicitly, in words, depict the group of individuals who had both a positive test result but not cancer (P[Po&¬Ca]) as a subset of the total individuals without cancer (P[¬Ca]). In the hard condition, they also had to express the group of individuals who had a positive result and cancer (P[Po&Ca]) as a subset of the individuals with cancer (P[Ca]). A mathematical formula was not sufficient to be assigned this code. An example comes from P261 who stated, "Of the 90% who do not have cancer, 15% will get a positive mammography". Here we can see a word-based representation of the individuals who do not have cancer but got a positive test result as a subset of those who do not have cancer. This classification was applied conservatively. For example, P498 who said "First what is 15% of 90%, that is 13.5%" did not receive the classification. An example from the hard condition which did get this

classification comes from P138 who said "We know 10% of women will have breast cancer in the screen and 80% of those will show up positive […] Of the remaining 90 women who do not have breast cancer 10% will be given a false positive so an additional 9 women."

A BLR showed that this representation was unsurprisingly more common within the PP (24.0%) condition, which expressed the problem in this format, than in the SP (9.7%) condition (Wald $X^2 = 18.0$, $p<.001$), which used an individual chance format. However clearly some individuals in the SP condition re-represented the problem in terms of nested sets. Furthermore, in both conditions, this representation was highly associated with normativity, as can be seen in Table 1.

A BLR was run with normativity as DV, and condition and NS-representation as IV's, and a unique predictive effect of NS-representation (Wald $X^2 = 123.6$, $p<.001$) was found, but no unique effect of condition (Wald $X^2 = 0.04$, $p=0.837$).

## Conversion to frequencies

Across all conditions, 87 participants (16.7%) also converted the base rate in the problem from a percentage into a frequency before attempting solution (i.e. before providing an NS-representation or completing the first computational step). For this classification, a 'sample' or 'population' of individuals as a frequency rather than a percentage or probability had to be expressed. For example, P105 said 'To make my math easier, I am going to assume there are 100 women.' and P186 began 'Out of 100 women, 10 have breast cancer, while 90 do not.' Out of the 87 participants who converted the problem to whole numbers, 73 converted to a population of 100 women. The number of individuals who made this conversion in each condition, crossed with those providing the NS-representation and the proportion of these subgroups providing the normative response can be seen in Table 1. A BLR with conversion as DV and condition as IV showed a predictive effect (Wald $X^2 = 7.3$, $p=.007$). A BLR with normative response as DV and condition and conversion as IV's showed a unique effect of conversion (Wald $X^2 = 128.9$, $p<.001$) and a potential unique effect of condition (Wald $X^2 = 5.2$, $p = 0.041$).

To simultaneously test the impact of condition, NS-representation and conversion upon normativity, a BLR was run. No main effect of condition was seen (Wald $X^2 = 0.172$, $p=0.68$), but a unique effect of NS-representation (Wald $X^2 = 93.2$, $p<.001$) and of conversion (Wald $X^2 = 8.3$, $p=0.004$) was seen. A table depicting these relationships can be seen below.

Table 1. Percentage of individuals providing the normative answer organized by condition, NS-representation and conversion (total number of individuals in each subgroup regardless of normativity in parentheses).

| | | Standard Probability | | | Proportion Percent | | |
|---|---|---|---|---|---|---|---|
| | | No NS-representation | NS-representation | Total | No NS-representation | NS-representation | Total |
| No-Conversion | | 1.4 (221) | 69.2 (13) | 5.1 (234) | 2.8 (176) | 45.8 (24) | 8.0 (200) |
| Conversion | | 5.0 (20) | 84.6 (13) | 36.4 (33) | 5.9 (17) | 78.4 (37) | 55.6 (54) |
| Total | | 1.7 (241) | 76.9 (26) | (267) | 3.1 (193) | 65.6 (61) | (254) |

From the raw data, we can see that in the absence of the NS-representation, conversion only appears to be associated with a small (~3%) increase in normativity, while in the presence of the NS-representation, converting appears to be associated with a much larger (~15-30%) increase. To check this, we ran two BLR's, predicting normativity from conversion. Within those who did not produce an NS-representation, no predictive relationship was seen (Wald $X^2 = 0.81$, $p=0.21$) while within those who did produce an NS-representation, a predictive relationship was seen (Wald $X^2 = 6.4$, $p=0.011$). Dependency of this sort was not seen for the NS-representation, which was a significant predictor of normativity among those who did not convert (Wald $X^2 = 69.3$ $p<.001$) as well as those who converted (Wald $X^2 = 27.6$, $p<.001$). For some individuals their process could not be determined (e.g. if they just provided a mathematical formula) but a few individuals were able to solve the problem without converting and also while apparently using a chance representation, such as P40:

*"There is a 10% chance that any woman over 40 has breast cancer [and] there is a 10% chance that a woman who does not have breast cancer over 40 gets a positive result. This means there is a 9% chance of [a false positive] and a 19% chance that someone tests positive for breast cancer. Out of this there is a 10/19% chance that the diagnosis is correct meaning there is a 52.63% chance."*

## Errors

The most common error within the SP condition (21.7%) was to provide the complement of the false positive rate, (1-P[Po|¬Ca]). This was much less common within the PP condition (5.5%). The TA data was coded for insight into common reasoning and a single piece of reasoning was highly prominent (45.8% of cases). This was the confusion of P(Po|¬Ca) with P(¬Ca|Po). Following this confusion, the subsequent accurate

deduction was made that 100% minus this value would give P(Ca|Po). For example, P228 said 'The fact that 15% of positive mammographies are invalid means that 85% are valid. She therefore has an 85% chance of actually having breast cancer', P20 said 'I guess since 10% of positive tests are inaccurate, that means there's a 90% chance of her having cancer' and P133 said 'Also of all the women who get a positive mammogram, 15% will not have breast cancer, so I think it is 85%.' Each of these participants use language reflecting P(¬Ca|Po) but accompanying the percentage value representing P(Po|¬Ca), strongly suggesting a confusion between the two. P177 expressed this confusion more explicitly, saying 'But there is a 10 percent chance that a woman without breast cancer will get a positive mammogram [true, P(Po|¬Ca)], so 10 percent of the positive mammograms are not accurate [false, P(¬Ca|Po)]'. In the remainder of these participants' TA data, the reasoning could not be extracted from the data. For example, many participants simply provided mathematical notation.

### Computational Steps

A cumulative graph depicting the proportion of individuals reporting each of the three computational steps, step 1 the calculation of P(Po&Ca) and P(Po&¬Ca), step 2 the summing of these and step 3 the division of P(Po&Ca) by the sum as well as whether the participant provided the normative numerical value can be seen below for both conditions. For both conditions, the majority of individuals do not achieve step 1, with further substantial but smaller drop-off between this and step 2, and no substantial subsequent drop-off between these and step 3 or the normative response. In short, highly similar curves were seen for both the SP and PP conditions. The major difference between the two conditions was the number of individuals reporting step 1 (with more individuals reporting this in the PP condition). Similar proportional drop-off was subsequently seen in both conditions. Indeed, while condition was predictive of step 1 (Wald $X^2$ = 15.3, p<.001), when controlling for step 1, condition was not predictive of step 2 (Wald $X^2$ = 0.19, p=.891), step 3 (Wald $X^2$ = .988, p=.320) or the normative response (Wald $X^2$ = 0.076, p=.783).

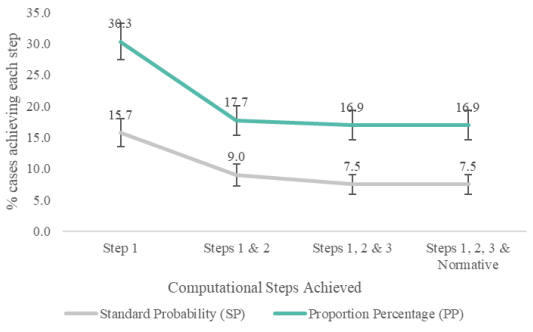


Figure 2. Drop-off graph for each computational step. Error bars represent one standard error.

## Discussion

We replicated Macchi's (2000) finding in a larger sample, and across a range of different format types, including with percentage base rates with and without the possibility of false negatives and with whole numbers and non-whole numbers. In each case, Macchi's proportion percentage format improved normativity over and above the SP format, with an overall increase from 9.0% to 18.1%.

We found that normativity is highly associated with the individual reporting a representation of P(Po&¬Ca) as a subset of P(¬Ca) in their think aloud data, and in the hard condition, also P(Po&Ca) as a subset of P(Ca). This finding is not surprising within the proportion percentage group, as it could be argued that these individuals are simply regurgitating the text from the problem. However, crucially, this relationship also held within the standard probability format, where an 'individual chance' probability format (i.e. 'If a woman has cancer, her chance of …) was presented. This observational finding should also be considered in the context of previous experiments (e.g. Evans et al, 2000; Sloman et al, 2003) showing that attempts to assist individuals in creating exactly this representation of the problem have been successful in increasing accuracy. Here we show that some individuals, without any prompt to do this, spontaneously adopt this representation, and this correlates highly with normativity. We also found some evidence that the NS-representation may have a mediating effect on the impact of the NS format. This provides some complementary evidence to those papers that the mechanism by which nested sets formats achieve greater accuracy is at least partially that which they have espoused: by encouraging a nested sets representation of the structure of the problem.

We also found that many individuals make a further spontaneous re-representation of the problem, and that this also correlates highly with normativity. This is the conversion of the problem from a percentage format into a frequency format. Interestingly, conversion alone seemed not to be predictive of normativity, however in combination with the NS-representation it was associated with higher rates of normativity than the NS-representation alone. The same was not true of the NS-representation. This was still highly predictive of normativity with or without conversion. Importantly, the majority of individuals who converted did so to a base of 100, making no mathematical change to the problem. This therefore seems to demonstrate a preference among our sample for working with frequency values over percentages, even when the absolute numbers (e.g. 20% vs 20 women out of 100) and therefore calculations, are identical. Of course, we cannot resolve the ultimate reason for this, be that a greater evolutionary exposure towards frequencies or a current greater exposure to frequencies during our participants' lives. We tentatively suggest a third

possibility. It may be difficult to mentally represent a percentage, abstract as it is, without it being a percentage of something tangible. Imagining 100 women may simply provide a concrete mental image which can be divided and sub-divided according to the percentages. It may also provide a platform for a simple internal narrative about these women and what happens to them. Whatever the ultimate reason however, this result does partially confirm Hoffrage et al's (2002) conjecture.

These findings have some relevance to the question of whether the elements that are thought to comprise the natural frequency format are separable, and if so, which elements are doing the 'work' in improving accuracy. Nested sets advocates have argued that the nested sets structure is doing all the work, and the frequencies are superfluous. Natural frequency advocates have argued that the two are inseparable. Here we find some tentative evidence that the two are separable (individuals who form a nested sets representation but do not convert to frequencies are still more successful than those who do not form that representation). However, even if separable, both the nested sets structure, and the use of frequencies (as opposed to percentages) appear to uniquely contribute to success, with the combination of both being more strongly associated with success than either alone. Importantly, without the nested sets structure, conversion to frequencies did not predict success, which may mirror findings that normalized frequency formats are no better than the standard probability format (e.g. Evans et al., 2000).

In terms of further investigation into the mechanisms of Macchi's nested sets format, we presented evidence that relative to the SP format, more individuals achieve step 1 (de-normalization). However, controlling for this, the proportion of participants achieving subsequent steps is not different to the SP format. Related to this, an analysis of errors between conditions has shown that the classic 1-P(Po|¬Ca) error was drastically reduced from 21.7% of total responses in the SP format to 5.5% in the PP format. This error, in line with previous theorizing (e.g. Braine and Connell, 1990) has been found here to principally stem from a confusion between the false positive rate P(Po|¬Ca) and P(¬Ca|Po). As has been mentioned, the clarification of the false positive rate (and the true positive rate in the hard condition) by encouraging individuals to see it as a subset of P(¬Ca) has long been theorized to be the mechanism by which nested sets formats work. The reduction of this error in the PP condition therefore seems to further support this theory. Given that the false positive rate is required for step 1, it also provides further evidence that the impact of Macchi's format is principally achieved at this step.

As noted, Macchi's format does not appear, upon the current evidence, to provide any additional support in the later stages of solution, most notably in getting from computational step 1 to step 2. At this step individuals need to recognize (A) that they require the total number of positive results, and (B) that they need to combine the false positives with the true positives to achieve this. So far, research has been principally focused on helping solvers form a representation of e.g. P(Po&¬Ca) as a subset of P(¬Ca). However, success on the final two steps may instead be a product of recognizing a different set relation, that of P(Ca&Po) and P(¬Ca&Po) as subsets of P(Po). We can clarify this distinction by displaying two tree structures of the medical diagnosis problem below. The top shows the classic structure, widely published, with the hypothesis, 'Cancer' as the first 'division', or first set of child nodes. However, the opposite structure is also possible, shown at the bottom, with the data, 'Positive' as the first set of child nodes.

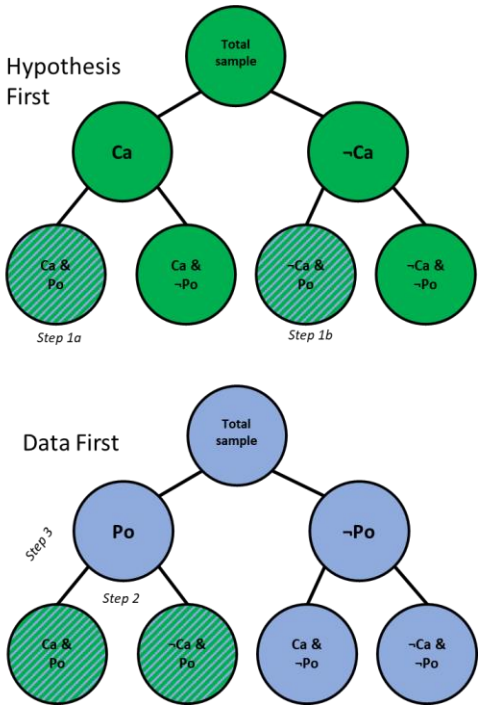


Figure 3. 'Hypothesis First' and 'Data First' tree diagram representations of the medical diagnosis problem.

While perceiving the set relations in the hypothesis-first version seems key to step 1, steps 2 (calculating total positives) and 3 (dividing cancer & positive by total positives) would seem to require an understanding of the set relations in (at least the left half) of the data-first diagram. For step 2, the addition, one must understand that P(Ca&Po) and P(¬Ca&Po) are subsets of P(Po). It seems to us that step 3, the division, should require only that same set relation i.e. that P(Ca&Po) is a subset of P(Po). To our knowledge this distinction has not been made before. We believe that in order to improve framing methods further, focus should be on helping individuals form these latter set representations at the most appropriate time to facilitate steps 2 and 3.

In the medical diagnosis problem, the information related to steps 2 and 3 are contained within the question. In

our nested sets format, this is changed into a proportion form i.e. *'What percentage of women at age forty who get a positive mammography…*', unlike the SP format, which is chance framed. While plausibly this could have helped solvers form exactly this latter subset representation, the current evidence suggests this did not have an impact. Future work may look to combine Macchi's format with a question form used by Girotto and Gonzalez (2001) which was divided into two parts: first explicitly requiring the calculation of step 2, and only then requiring calculation of step 3.

Finally, it should be noted, that the accuracy percentage for participants in our NS group was lower than the average from the recent meta-analysis for natural frequency (~24%). It is difficult of course to make confident comparisons but given that we have found that the nested sets approach works via very similar mechanisms to the natural frequency approach, but requires one extra step (de-normalization), and in some versions two extra steps, and furthermore that we have found a unique beneficial effect of frequencies, some greater accuracy on natural frequency versions seems plausible to us. Pragmatically therefore we would still advocate for natural frequencies as the primary method for communicating Bayesian problems to the public where that is possible, with proportion percentages as a backup where it is not.

However, unfortunately when individuals do encounter Bayesian problems in the real world, they are often in the standard probability format. Sedlmeier & Gigerenzer (2001) have investigated the merits of preparing individuals via training to convert these into natural frequency versions themselves when they encounter them. This however requires considerable training. Our findings suggest that solvers can do more of the work themselves than was assumed by that research (i.e. can de-normalize the problem themselves) and therefore may only need to remember fewer 'conversion' steps. This may be valuable where the brevity of the training is important. In fact, our findings tentatively suggest substantial accuracy gains may be obtained by training people to following two simple rules when faced with an SP problem:

1. Imagine 100 women (or whatever unit you're dealing with).
2. Imagine the percentages you've been given as proportions of these 100 women.

## Acknowledgements

Funding was in part provided by the ERC project ERC-2013-AdG339182-BAYES_KNOWLEDGE and the Leverhulme Trust project RPG-2016-118 CAUSAL-DYNAMICS.

## References

Brase, G., & Hill, W. (2015). Good fences make for good neighbors but bad science: a review of what improves Bayesian reasoning and why. *Frontiers in Psychology*, 6, 340.

Braine, M & Connell, J. (1990). Is the base rate fallacy an instance of asserting the consequent?. *Lines of thinking: Reflections on the psychology of thought*, 1, 165-180.

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

Gigerenzer, G. & Hoffrage, U. (1995). How to Improve Bayesian Reasoning Without Instruction: Frequency Formats. *Psychological Review*, 102(4), 684–704.

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

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

Johnson, E. & Tubau, E. (2013). Words, numbers, & numeracy: Diminishing individual differences in Bayesian reasoning. *Learning and Individual Differences*, 28, 34–40.

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.

McDowell, M., & Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychological Bulletin*, 143(12), 1273–1312.

McNair, S. (2015). Beyond the status-quo: research on Bayesian reasoning must develop in both theory and method. *Frontiers in Psychology*, 6, 1–3.

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

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