SOM-R: MATHEMATICAL PROOF AND LIKELIHOODS EXPERIMENT

A: Mathematical Proof Illustrating Irrelevance of Priors

For the following model:

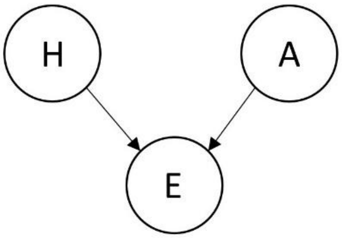

Figure A1. Simple common-effect structure. Hypotheses H and A are non-exhaustive and non-exclusive possible causes of evidence E.

Let:

$$P(E|H,A) = x$$

$$P(E|H,\neg A) = y$$

$$P(E|\neg H,A) = z$$

$$P(E|\neg H,\neg A) = w$$

And suppose

$$x > y$$

$$x > z$$

$$y > w$$

$$z > w$$

We will assume $P(H) = h$ and $P(A) = a$, and that $0 < h < 1$ and $0 < a < 1$.

Then we have to prove $P(H|E) > P(H)$.

Proof

First note that by Bayes,

$$P(H|E) = P(E|H)P(H)/P(E)$$

So we have to show

$$P(E|H)P(H)/P(E) > P(H)$$

However, this is the same as showing

$$P(E|H)/P(E) > 1$$

Which in turn, is the same as showing

$$P(E|H) > P(E)$$

Consequently, to prove $P(E|H) > P(E)$, we consider the 2 cases: When A is true, and when A is false.

Case 1: When A is true

In this case

$$P(E|H) = P(E|H,A) = x$$

By marginalisation:

$$\begin{aligned} P(E) &= P(E|H,A)P(H) + P(E|\neg H,A)P(\neg H) \\ &= xh + z(1-h) \end{aligned}$$

So

$$\begin{aligned} P(E|H) - P(E) &= x - xh - z(1-h) \\ &= x(1- h) - z(1-h) \\ &= (x- z)(1-h) \end{aligned}$$

And since $x > z$,

$$P(E|H) - P(E) > 1-h$$

And therefore since $0 < h < 1$,

$$P(E|H) - P(E) > 0$$

This completes the proof for case 1.

Case 2: When A is false

In this case

$$P(E|H) = P(E|H,\neg A) = y$$

By marginalisation:

$$P(E) = P(E|H,\neg A)P(H) + P(E|\neg H,\neg A)P(\neg H)$$

$$= yh + w(1-h)$$

So

$$P(E|H) - P(E) = y - yh - w(1-h)$$

$$= y(1- h) - w(1-h)$$

$$= (y- w)(1-h)$$

And since $y > w$,

$$P(E|H) - P(E) > 1-h$$

And therefore since $0 < h < 1$,

$$P(E|H) - P(E) > 0$$

This completes the proof for case 2.

And hence we have proven $P(E|H) > P(E)$ for both cases.

E: Experiment 4: Manipulating Likelihood Ratios

Experiment 4 was designed to determine whether the zero-sum fallacy holds even when the likelihoods are not identical. In this way, it was possible to determine whether the fallacy is shallow – requiring the likelihoods to be exactly equal – or whether the presence of unequal (but still predictive) likelihoods are sufficient to induce the fallacy. Importantly, from a normative (Bayesian) standpoint, support for both hypotheses should increase given a positive test – irrespective of the specific likelihood ratios – as long as the general inequalities hold (i.e. $P(E|\neg H_1,\neg H_2) < P(E|H_1,\neg H_2)$ and $P(E|\neg H_1,H_2) < P(E|H_1,H_2)$).

Method

Participants. Participants were recruited using the same protocol as in Experiment 1. A sample size of approximately 200 was predetermined, based on the same rationale as Experiments 2 and 3. Of the 204 participants recruited, three were removed whose native language was not English or were living outside the US. Of the 201 participants remaining, 97 were female. The mean age was 35.92 (*SD* = 11.55). Participants were paid $1 for their time (*Median* = 7.13 minutes, *SD* = 4.83).

Materials and Procedure. The materials and general procedure generally followed that of Experiment 2, barring the following exceptions:

Along with the between-subject factor of test result condition (positive or negative; common to Experiments 1, 2, and 3), an additional, 2-level between-subject "likelihoods" factor was added (making a 2x2 between-subject design). This factor consisted of two possible conditions, based on the direction of inequality between the two test likelihoods. In one condition ("Main Hypothesis Favoured"; MHF) the likelihoods always favoured the main hypothesis (i.e. a positive test result was more likely given the hypothesis in question – e.g. using steroids – than the alternative hypothesis). This inequality was reversed in the other condition ("Main Hypothesis Disfavoured"; MHD). Although for each condition all four scenarios had inequalities in the same direction, the degree of inequality varied across scenarios (differences of 10%, 8%, 6%, and 5%; see Supplementary Materials F for example scenarios). Degrees of inequality were fixed to a given scenario context across all conditions and were always equal across likelihood manipulation.

Accordingly, as participants completed the 4 scenarios in a random order, they made a judgment, expressed their confidence in that judgment, and then provided some reasoning.

Results

Judgment Data. Each of the 201 participants made 4 judgments, resulting in a total of 804 judgments. Fig. B.1 shows the mean proportions of these judgments, split by test result condition (columns) and likelihoods manipulation (rows).

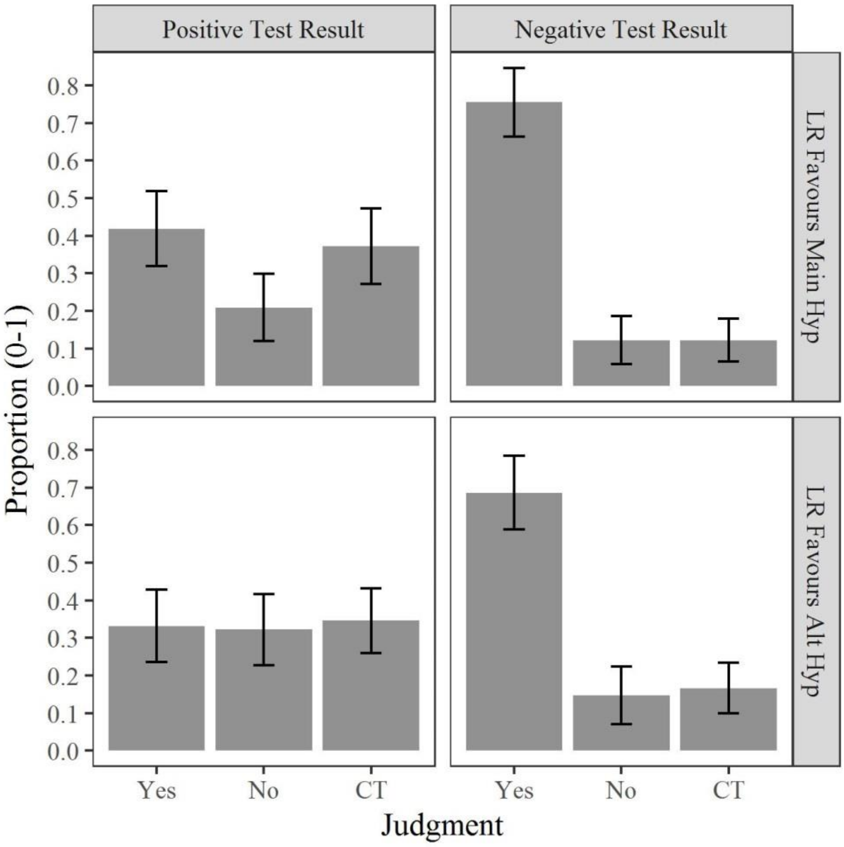

Fig. B.1

Experiment 4. Mean proportions of judgments (CT = “Cannot Tell” responses), split by test result condition (columns) and likelihoods manipulation (rows). Error bars reflect 95% Confidence Intervals.

Following the analysis protocol of the preceding experiments, participant judgments were again coded into a single, summary correct responding variable, which was used as the dependent variable in subsequent Bayesian ANOVA. Hierarchical model comparison found that whilst correct responding was significantly higher in negative (vs positive) test result conditions (right vs left columns of Fig. B.1), $BF_{Inclusion} = 2.75 * 10^9$, there was no effect of likelihoods manipulation, $BF_{Inclusion} = 0.48$. Consequently, the model with only a main effect of test result was both the best fit, $BF_M = 5.57$, and significant overall, $BF_{10} = 3.5 * 10^9$.

Table B.1.

Experiment 4: Correct responding descriptives and chance responding analysis, split by condition.

Test Result	Likelihood Manipulation	*M*	*SD*	*N*	*≠1.33 (BF₁₀)*	*δ*	*δ 95% CI*
Negative	Main Hyp Disfavored	2.75	1.34	51	$1.316 * 10^7$	1.02	0.704, 1.365
	Main Hyp Favored	3.02	1.22	49	$1.332 * 10^{10}$	1.346	0.971, 1.732
Positive	Main Hyp Disfavored	1.33	1.32	52	0.151‡	-0.005	-0.275, -0.259
	Main Hyp Favored	1.67	1.33	49	0.701†	0.245	-0.030, 0.529

Note: *† = anecdotal evidence, ‡ = strong evidence for the null.*

In line with previous experiments, correct responding in negative test result conditions was significantly greater than chance (right-hand column of Table B.1), whilst correct responding in positive test result conditions was at chance level.

Lastly, as in Experiments 1, 2, and 3, using Bayesian contingency tables, judgments were shown to be unaffected by scenario order, $BF_{10} = 2.24 * 10^{-4}$, and scenario type, $BF_{10} = .001$, with very strong evidence for the null in both cases.

Confidence Data. Fig. B.2 shows the boxplot breakdown of confidence by judgment type (within-pane), test result condition (columns) and likelihoods manipulation (rows).

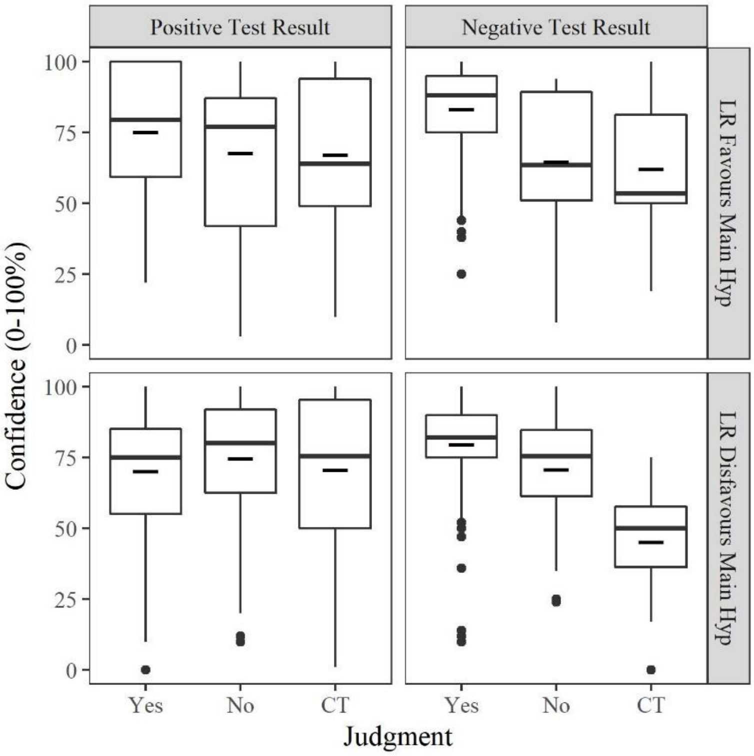

Fig. B.2

Experiment 4. Confidence in judgments (CT = “Cannot Tell” responses), split by test result condition (columns) and likelihoods manipulation (rows). Whiskers represent +/- 1.5 * IQR. Means shown as short crossbars.

A Bayesian ANOVA was run on the three variables of interest, Judgment (3) x Test Result Condition (2) x Likelihoods Manipulation (2). Hierarchical model comparisons revealed significant main effect of judgment on confidence, $BF_{Inclusion} = 2.03 * 10^{11}$, with “Cannot Tell” responses as least confident, and “Yes” responses as most confident. There was no main effect of test result condition, $BF_{Inclusion} = 2.45$, and strong evidence for a null effect of likelihood manipulation, $BF_{Inclusion} = 0.23$. Lastly, the analysis yielded a significant interaction between test result condition and judgments, $BF_{Inclusion} = 3.95 * 10^{6}$, with the model including this interaction term yielding the most significant model improvement, $BF_{M} = 3.95 * 10^{6}$, and decisive evidence overall, $BF_{10} = 2.16 * 10^{17}$. To explore this interaction further, a second round of ANOVA were performed on both positive and negative test result conditions in isolation. This revealed that confidence was not significantly affected by judgment in the positive test result condition (see left-hand column of Fig. B.2; N = 404), BF_{10} = 0.08, whilst those in the negative test condition were *decisively* less confident in “Cannot Tell” and “No” judgments than “Yes” judgments (see right-hand column of Fig. B.2; N = 400), BF_{10} = 4.98 * 10^{23}.

Discussion

The zero-sum fallacy was found in Experiment 4, replicating Experiments 1, 2 and 3. Further, such effects do not rely on superficial reasoning regarding the equality of likelihood values, indicating a more fundamental misapprehension about the probative value of the test result. Lastly, the confidence findings of Experiments 2, and 3 were also replicated, further supporting the existence of a genuine reasoning bias.