**Supplementary material 3 – Analysis of the experiments**

**Code for deriving posteriors in Experiment 2**

# priorn1, priorn2 = vague prior on the number of patients from subpopulations S1 and S2,

# respectively, treated with the pill (provided to participants)

# priors1, priors2 = vague prior on the number of patients from subpopulations S1 and S2,

# respectively, whose symptoms resolved after treatment (provided to participants)

# nsamp1, nsamp2 = number of patients from subpopulations S1 and S2, respectively, treated on

# each clinic day (observed in the game)

# cures1, cures2 = number of patients from subpopulations S1 and S2, respectively, whose

# symptoms resolved after treatment, on each clinic day (observed in the game)

ps = proportion of observed patients who were from subpopulation S1

a1 <- 0.5 + priors1 + sum(cures1)

a2 <- 0.5 + priors2 + sum(nums(cures2)

as <- 0.5 + sum(nsamp1)

b1 <- 0.5 + sum(nsamp1) - priors1 + priorn1 - sum(cures1)

b2 <- 0.5 + sum(nsamp2) - priors2 + priorn2 - sum(cures2)

bs <- 0.5 + sum(nsamp2)

nsam <- 100000

p1 <- rbeta(nsam, a1, b1)

p2 <- rbeta(nsam, a2, b2)

ps <- rbeta(nsam, as, bs)

pel1 <- p1\*ps + p2\*(1 - ps) # posterior without extrapolation

pel2 <- p1\*0.5 + p2\*0.5 # posterior with extrapolation

**Table 1A: Sample description**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | N | Mean age (sd) | Male (%) | UG (%) | Year of study | Percentage with qualifications in quantitative subjects (n) | | Percentage confident in using probabilities (n) | | | | | BNT score,  out of 7 (sd) | Score on Scott and Bruce’s Decision Style Inventory, out of 5 (sd) | | | | |
| A level | AS level | 1 | 2 | 3 | 4 | 5 | R | I | D | A | S |
| Nursing | 18  (25%) | 23.1  (7.3) | 33.3  (6) | 83.3  (15) | 1.9  (0.8) | 11.1 (2) | 0 (0) | 5.6 (1) | 33.3 (6) | 38.9 (7) | 16.7 (3) | 5.6 (1) | 3.1 (1.6) | 4 (0.6) | 3.7 (0.5) | 3.7 (1) | 2.7 (0.9) | 2.7 (0.9) |
| Midwifery | 11  (15.3%) | 21.5  (5.6) | 0  (0) | 100  (11) | 2.1  (1) | 18.2 (2) | 0 (0) | 0 (0) | 45.5 (5) | 18.2 (2) | 18.2 (2) | 18.2 (2) | 3.5 (1.4) | 3.9 (0.6) | 3.5 (0.5) | 3.4 (1) | 2.4 (1.1) | 2.6 (0.8) |
| Medicine | 24  (33.3%) | 20.2  (3.3) | 8.3  (2) | 91.7  (22) | 1.2  (0.8) | 58.3 (14) | 12.5 (3) | 0 (0) | 45.8 (11) | 25 (6) | 25 (6) | 4.2 (1) | 4 (1.6) | 4.2 (0.4) | 3.5 (0.6) | 3.8 (0.6) | 2.3 (0.9) | 2.4 (0.8) |
| MSc in HE | 6  (8.3%) | 26.3  (3.5) | 33.3  (2) | 0  (0) | 1  (0) | 16.7 (1) | 66.7 (4) | 16.7 (1) | 50 (3) | 33.3 (2) | 0 (0) | 0 (0) | 3.8 (1.2) | 4.4 (0.5) | 3.2 (0.8) | 3.3 (0.6) | 2.4 (1) | 2.8 (0.7) |
| MSc in HS | 3  (4.2%) | 27  (7.2) | 0  (0) | 0  (0) | 1  (0) | 33.3 (1) | 0 (0) | 0 (0) | 100 (3) | 0 (0) | 0 (0) | 0 (0) | 3.3 (2.5) | 4.1 (1) | 3.2 (0.8) | 4.6 (0.7) | 2.7 (2) | 2.1 (0.3) |
| Psychology | 1  (1.4%) | 20  (N/A) | 0  (0) | 100  (1) | 2  (N/A) | 100 (1) | 0 (0) | 0 (0) | 100 (1) | 0 (0) | 0 (0) | 0 (0) | 4 (N/A) | 4 (N/A) | 4 (N/A) | 4 (N/A) | 2.8(N/A) | 1.6(N/A) |
| Other | 9  (12.5%) | 23.2  (6.7) | 44.4  (4) | 100  (9) | 1.8  (0.7) | 77.8 (7) | 0 (0) | 11.1 (1) | 44.4 (4) | 33.3 (3) | 11.1 (1) | 0 (0) | 4.4 (1.5) | 4.1 (0.5) | 3.5 (0.8) | 4 (0.6) | 2.5 (1) | 2.5 (0.6) |
| Total | 72  (100%) | 22.3  (5.7) | 19.4  (14) | 80.6  (58) | 1.6  (0.8) | 38.9 (28) | 9.7 (7) | 4.2 (3) | 45.8 (33) | 27.8 (20) | 16.7 (12) | 5.6 (4) | 3.7 (1.6) | 4.1 (0.5) | 3.5 (0.6) | 3.7 (0.8) | 2.5 (1) | 2.5 (0.8) |

HE=health economics, HS= health sciences. Students in health economics were only selected if their previous degree was healthcare related (medicine, pharmacy or dentistry). Other includes law, social policy, history, finance, mathematics, physics and economics

1=very confident; 2=somewhat confident; 3=neither confident nor unconfident; 4=somewhat unconfident; 5=very unconfident.

R=rational; I=intuitive; D=dependent; A=avoidant; S=spontaneous.

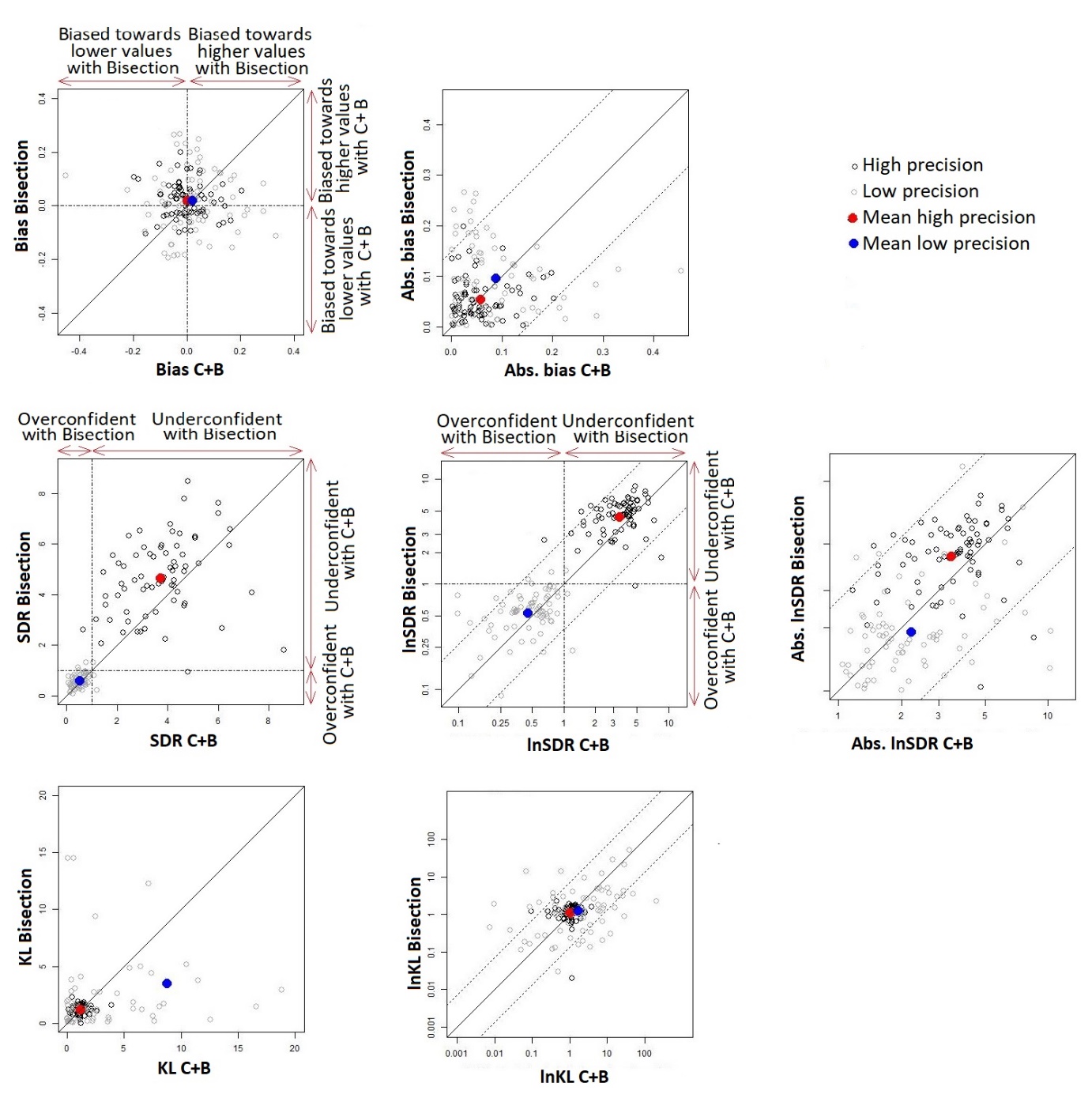
**Further analyses for experiment 1**

**Figure 1A. Distribution of bias, uncertainty and KL scores in Experiment 1. (SDR= SD elicited / SD true)**

|  |  |  |
| --- | --- | --- |
|  | SD  Ratio  (SD elicited / SD true) | **High precision Low precision** |
|  | LnSDR |  |

**Empirical summaries**

Participants’ accuracy for different scenarios and elicitation methods is summarised (for each metric of accuracy) in 2A. The values in the table summarise the data shown in 2A, and support the findings presented in Chapter 8.



g)

f)

e)

d)

c)

a)

b)

**Figure 2A. Within-participant comparison of accuracy when different elicitation methods were used, for each precision scenario. Axis labels on d), e) and g) are on natural scale.**

Points below the diagonal line suggest participants were less accurate (higher KL) when using Chips and Bins.

**Table 2A. Aggregate comparison of accuracy when different elicitation methods were used, for each precision scenario.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Outcome  Mean(sd)  [median (IQR)] | Low precision  (high heterogeneity, 10 clinic days) | | High precision  (no heterogeneity, 25 clinic days) | |
| Bisection | C+B | Bisection | C+B |
| Bias (diff. in means) | 0.018 (0.118)  [0.011 (-0.058 to 0.096)] | 0.018 (0.122)  [0.012 (-0.051 to 0.065)] | 0.019 (0.066)  [0.021 (-0.03 to 0.06)] | 0 (0.076)  [-0.006 (-0.039 to 0.048)] |
| Absolute bias | 0.096 (0.071)  [0.086 (0.033 to 0.146)] | 0.088 (0.085)  [0.059 (0.034 to 0.119)] | 0.055 (0.041)  [0.041  (0.026 to 0.073)] | 0.057 (0.049)  [0.043 (0.022 to 0.076)] |
| SDR | 0.591 (0.246)  [0.553  (0.443 to 0.733)] | 0.518 (0.243)  [0.515 (0.339 to 0.683)] | 4.658 (1.512)  [4.59  (3.637 to 5.618)] | 3.715 (1.479)  [3.773 (2.752 to 4.518)] |
| lnSDR | -0.625 (0.485)  [-0.593 (-0.814 to -0.311)] | -0.79 (0.563)  [-0.664 (-1.081 to -0.382)] | 1.476 (0.38)  [1.524 (1.291 to 1.726)] | 1.223 (0.454)  [1.328 (1.012 to 1.508)] |
| Absolute lnSDR | 0.643 (0.46)  [0.593 (0.311 to 0.814)] | 0.799 (0.549)  [0.664 (0.382 to 1.081)] | 1.477 (0.376)  [1.524 (1.291 to 1.726)] | 1.235 (0.42)  [1.328 (1.012 to 1.508)] |
| KL | 3.536 (7.472)  [1.264 (0.44 to 2.97)] | 8.775 (26.027)  [2.102 (0.486 to 7.198)] | 1.218 (0.374)  [1.221 (1.031 to 1.456)] | 1.151 (0.564)  [1.104 (0.874 to 1.323)] |
| lnKL | 0.233 (1.433)  [0.234 (-0.823 to 1.088)] | 0.509 (2.037)  [0.741 (-0.722 to 1.974)] | 0.109 (0.575)  [0.199 (0.03 to 0.376)] | 0.019 (0.542)  [0.099 (-0.134 to 0.28)] |

**Modelling**

*Methods*

The different outcomes, y, were modelled using distributions whose mean E(y) is related to the various explanatory variables as follows:

|  |  |
| --- | --- |
| E(yij) = g-1(mu + pi + qprobij + b1\*Xij + b2\*Uij + b3\*Uij\*Xij). | (Equation 1) |

where j is the subject, i is the period, pi is the fixed effect of the ith period (where p1 = 0), qprobij the fixed effect of the probability value used by subject j in period i where probij assumes values {0,1,2,3} for values of probability of {0.3, 0.4, 0.6, 0.7}, Xij identifies the method of elicitation and is 0 if chips and bins is used by subject j in period i, or 1 if bisection is used, Uij identifies the precision scenario taking a value of 0 if subject j in period i completes an experiment for the low precision scenario, and 1 otherwise, and g() is a link function.

Bias is modelled using a Normal distribution with identity link. KL and SDRs are non-negative, and two model specifications were explored: 1) Normal distribution with identity link (where the outcome is log transformed); and 2) Gamma distribution with log link.

In addition, an alternative specification for all models was evaluated where a term on sj was added to describe subject-specific random effects (on the intercept), assumed to be normally distributed on the appropriate scale, with mean zero and an estimated standard deviation. The original specification where subject is omitted is identified as fixed effect model throughout.

To compare the model fit between the above-mentioned specifications, the Akaike Information Criterion (AIC) was used. AIC for the Gamma model was adjusted by adding to make it comparable to that of the Gaussian model.

*Results*

Estimates of the model coefficients and their standard errors are shown in Table 3A, alongside the standard deviation of the random effect subject term (when applicable), and model AIC. AICs favoured the fixed effect model for bias and KL divergence, and the random effects model for SDR, indicating a lack of within-person correlations in responses for bias and KL, and weak within-person correlations for SDR.

Results suggest that mean bias is low or null, and show no evidence that any coefficient is different to zero (wide standard errors in relation to the mean). For SD ratios, AIC suggests better fit for the Gamma model than the Gaussian, but the coefficients are comparable, and consistent. The coefficients suggest that participants were more certain when using Chips and Bins than bisection, and less certain in high precision scenarios. For KL divergence, AIC suggests better fit for the Gamma model.

The fitted models also enable comparison of estimates and confidence intervals for expected outcomes between groups, in 4A.

**Table 3A. Regression model coefficients for experiment 1. Note that intercepts are omitted from the table as they are not comparable between Gamma and Gaussian models**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Bias** | | **Ln SD ratio** | | | | **Ln KL** | | | |
|  | **Gaussian** | | **Gaussian** | | **Gamma** | | **Gaussian** | | **Gamma** | |
|  | FE | RE | FE | RE | FE | RE | FE | RE | FE | RE |
| **Coefficients** |  |  |  |  |  |  |  |  |  |  |
| **C+B** | -0.003  (0.015) | -0.003  (0.014) | -0.161 (0.079) | -0.161 (0.061) | -0.131 (0.067) | DNC | 0.268  (0.218) | 0.268  (0.211) | 0.808 (0.258) | 0.721 (0.168) |
| **High precision** | -0.003  (0.015) | -0.003  (0.014) | 2.105 (0.079) | 2.105 (0.061) | 2.065 (0.068) | DNC | -0.127  (0.218) | -0.127  (0.211) | -1.015 (0.258) | -0.539 (0.174) |
| **C+B High precision** | -0.014  (0.021) | -0.014  (0.02) | -0.097 (0.112) | -0.097 (0.086) | -0.103 (0.095) | DNC | -0.353  (0.309) | -0.353  (0.298) | -0.853 (0.365) | -0.784 (0.229) |
| **Standard deviation for RE on subject** | -- | 0.028 | -- | 0.562 | -- | DNC | -- | 0.147 | -- |  |
| **AIC** | -559.1 | -491.1 | 400.4 | 386.5 | 367.4 | DNC | 983.6 | 1000.4 | 994.8 | 969.0 |

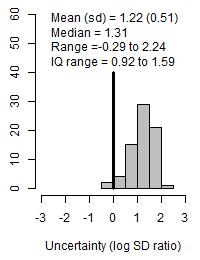
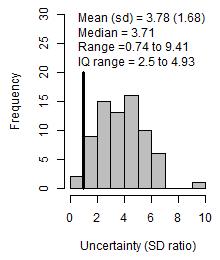
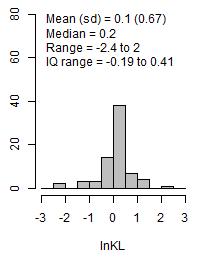
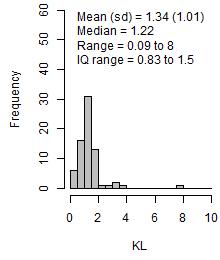
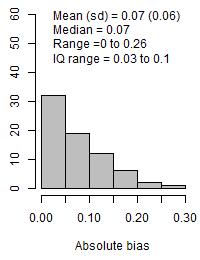
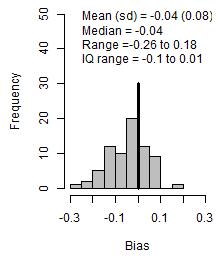
\*DNC: Did not converge

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Estimate [confidence interval] | Low precision  (high heterogeneity, 10 clinic days) | | High precision  (no heterogeneity, 25 clinic days) | |
| Bisection | C+B | Bisection | C+B |
| **Bias, Gauss (FE)** | 0.05 [0.02, 0.08] | 0.05 [0.02, 0.08] | 0.05 [0.02, 0.08] | 0.03 [-0.002, 0.06] |
| **SDR Gauss (RE)** | 0.55 [0.47, 0.64] | 0.47 [0.40, 0.55] | 4.50 [3.86, 5.25] | 3.48 [2.99, 4.05] |
| **SDR Gamma (FE)** | 0.58 [0.50, 0.67] | 0.51 [0.44, 0.59] | 4.55 [3.93, 5.28] | 3.60 [3.10, 4.18] |
| **lnKL Gauss (FE)** | 1.01 [0.53, 1.49] | 1.28 [0.79, 1.76] | 0.88 [0.41, 1.36] | 0.80 [0.31, 1.28] |
| **lnKL Gamma (RE)** | 0.65 [0.27, 1.04] | 1.37 [1.01, 1.74] | 0.11 [-0.25, 0.48] | 0.05 [-0.34, 0.44] |
| **KL (transformed from lnKL scale),**  **Gauss (FE)** | 2.74 [1.69, 4.43] | 3.58 [2.21, 5.81] | 2.41 [1.50, 3.88] | 2.22 [1.37, 3.60] |
| **KL (transformed from lnKL scale),**  **Gamma (RE)** | 1.92 [1.31, 2.83] | 3.95 [2.75, 5.69] | 1.12 [0.78, 1.61] | 1.05 [0.72, 1.55] |

**Table 4A. Expected outcomes by precision scenario and method, evaluated from the different models.**

**Descriptive summaries from experiment 2**

Figure 3A shows distribution summaries for the accuracy metrics without extrapolation. Bias was positive but low. SD ratios indicate that participants were more likely to be underconfident, which is consistent with findings from experiment 1 given that participants here played a high precision game.

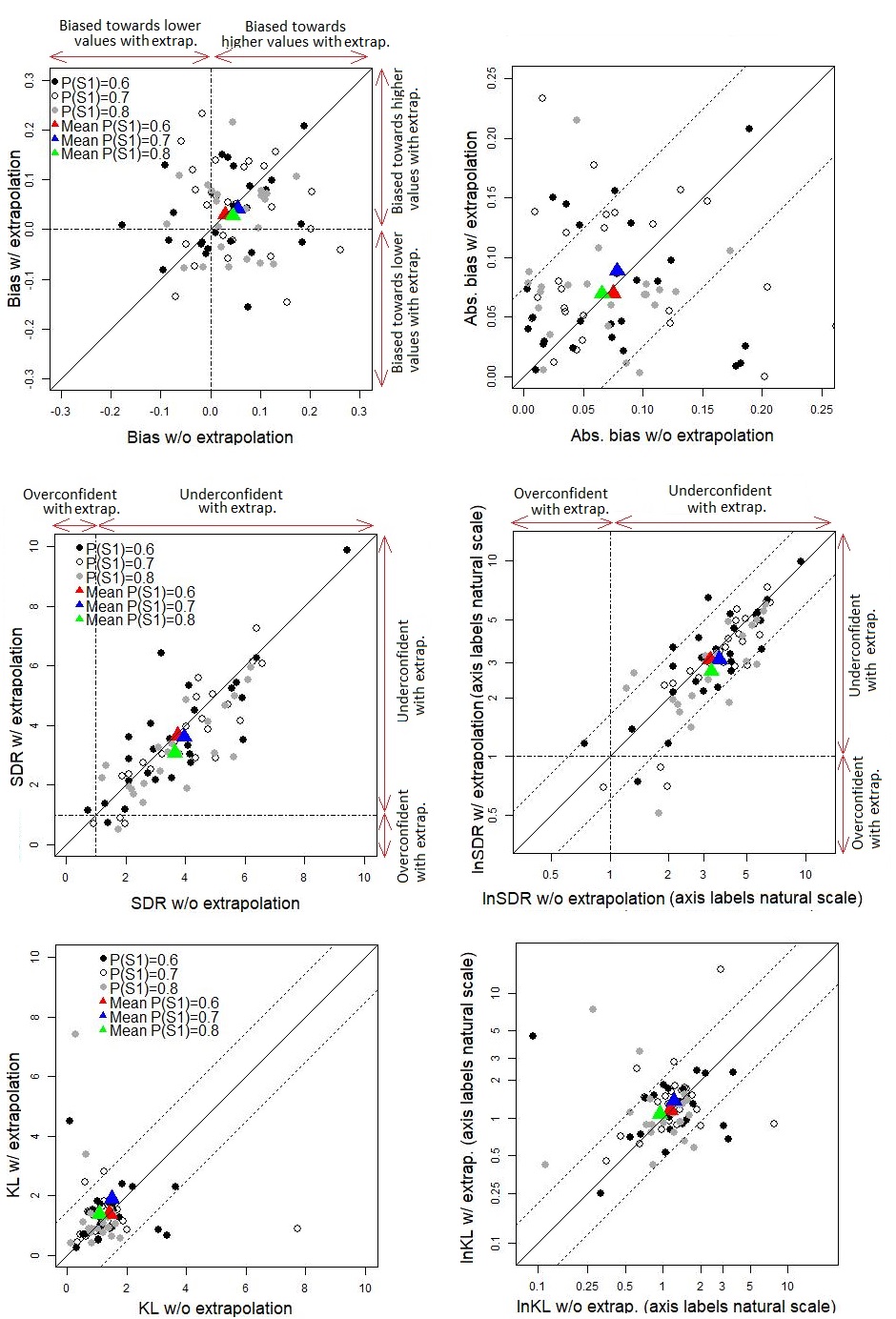


**Figure 3A. Distribution of bias, precision and KL scores, priors elicited without extrapolation.**

**Further analyses and results for experiment 2**

**Empirical summaries**

Participants’ accuracy in different scenarios and elicitation methods is summarised in Table 5A. The values in the table summarise the data shown in 4A, and support the findings presented in Chapter 8.



f)

e)

c)

d)

b)

a)

**Figure 4A. Within-participant comparison of accuracy with and without extrapolation, for different levels of extrapolation.**

**Table 5A. Aggregate comparison of accuracy with and without extrapolation, for different levels of extrapolation.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mean(sd)  [median (IQR)] | Split in S1 of 60% | | Split in S1 of 70% | | Split in S1 of 80% | |
| w/o extrapolation | w extrapolation | w/o extrapolation | w extrapolation | w/o extrapolation | w extrapolation |
| Bias | 0.03 (0.092)  [0.036 (-0.017 to 0.079)] | 0.055 (0.09)  [0.035 (-0.016 to 0.122)] | 0.046 (0.068)  [0.041 (0.006 to 0.102)] | 0.029 (0.085)  [0.011 (-0.027 to 0.086)] | 0.042 (0.098)  [0.051 (-0.03 to 0.125)] | 0.027 (0.078)  [0.05 (-0.054 to 0.076)] |
| Absolute bias | 0.075 (0.059)  [0.074 (0.025 to 0.095)] | 0.079 (0.069)  [0.05 (0.032 to 0.122)] | 0.066 (0.047)  [0.067 (0.018 to 0.102)] | 0.07 (0.055)  [0.047 (0.027 to 0.098)] | 0.088 (0.058)  [0.073 (0.049 to 0.136)] | 0.069 (0.043)  [0.071 (0.058 to 0.078)] |
| SDR | 3.728 (1.954)  [3.473 (2.112 to 4.306)] | 3.965 (1.537)  [4.025 (2.837 to 4.906)] | 3.631 (1.553)  [3.469 (2.347 to 4.937)] | 3.648 (2.048)  [3.314 (2.243 to 4.906)] | 3.602 (1.672)  [3.442 (2.732 to 4.684)] | 3.069 (1.426)  [2.988 (1.965 to 3.942)] |
| lnSDR | 1.174 (0.571)  [1.245 (0.748 to 1.46)] | 1.285 (0.476)  [1.392 (1.043 to 1.591)] | 1.189 (0.48)  [1.244 (0.851 to 1.597)] | 1.134 (0.606)  [1.198 (0.808 to 1.591)] | 1.137 (0.619)  [1.236 (1.005 to 1.544)] | 0.998 (0.551)  [1.094 (0.675 to 1.368)] |
| KL | 1.418 (0.871)  [1.151 (1.014 to 1.551)] | 1.503 (1.401)  [1.238 (0.97 to 1.514)] | 1.079 (0.45)  [1.217 (0.795 to 1.418)] | 1.373 (0.872)  [1.288 (0.81 to 1.69)] | 1.895 (2.842)  [1.339 (0.896 to 1.659)] | 1.378 (1.476)  [0.988 (0.794 to 1.298)] |
| lnKL | 0.137 (0.761)  [0.141 (0.014 to 0.439)] | 0.197 (0.603)  [0.213 (-0.031 to 0.415)] | -0.064 (0.644)  [0.196 (-0.229 to 0.349)] | 0.147 (0.606)  [0.253 (-0.211 to 0.525)] | 0.307 (0.656)  [0.292 (-0.11 to 0.506)] | 0.069 (0.625)  [-0.014 (-0.232 to 0.261)] |

**Modelling**

*Methods*

Each outcome y1 for the distribution elicited with extrapolation was modelled using a distribution with a mean defined by:

|  |  |
| --- | --- |
| E(y1j) = g-1(mu + b1\*y0j + b2\*I(Oj==70) + b3\*I(Oj==80) + b4\*Xj | (Equation 2) |

where j is the subject, y0j is individual j’s accuracy of the distribution elicited without extrapolation, Oj is the split individual j was randomised to, I() the indicator function that assumes a value of 1 if the enclosed statement is true and 0 if false, Xj the method of elicitation used by subject j (0 if bisection was used or 1 if chips and bins was used), and g() is a link function.

The model specifications explored were identical to experiment 1, as was the model selection process.

*Results*

The model coefficients are shown in 6A. AICs favour the Gaussian model for KL divergence, and the Gamma model for SD ratios and absolute bias. However, the results from different model specifications were consistent in all outcomes, in that extrapolation and the extent of extrapolation was not shown to have an effect on any of the outcomes (standard error of the relevant coefficients wide in relation to the mean).

The fitted models also enable us to compare estimates and confidence intervals for expected outcomes between groups, in 7A.

**Table 6A. Regression model coefficients for experiment 2, showing the effect of initial accuracy (without extrapolation), the extent of extrapolation, and the elicitation method on accuracy with extrapolation. Note that intercepts are omitted from the table, as they are not comparable between Gamma and Gaussian models.**

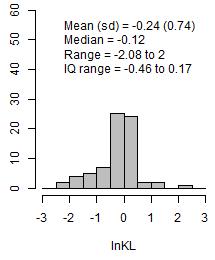
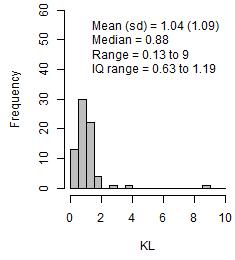
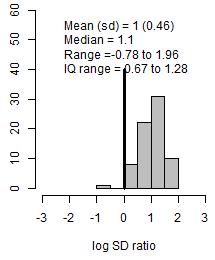
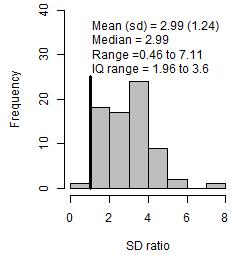
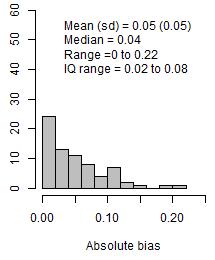
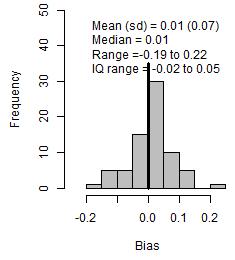
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Coef, mean (SE)** | **KL** | | **SD ratio** | | **Absolute bias (difference)** | | **Bias** |
| **Gaussian model**  **(log transformed)** | **Gamma model**  **(log link)** | **Gaussian model**  **(log transformed)** | **Gamma model**  **(log link)** | **Gaussian model**  **(log transformed)** | **Gamma model**  **(log link)** | **Gaussian model (natural scale)** |
| **Accuracy w/o extrapolation** | 0.102 (0.118) | 0.014 (0.177) | 0.931\* (0.086) | 0.858\* (0.081) | -0.111 (0.154) | 0.033 (0.079) | 0.05 (0.128) |
| **C+B** | 0.159 (0.179) | 0.302 (0.27) | -0.101 (0.101) | -0.108 (0.095) | 0.096 (0.387) | 0.267 (0.199) | 0.012 (0.025) |
| **Split S1 70%** | -0.06 (0.187) | -0.057 (0.281) | -0.149 (0.104) | -0.134 (0.098) | 0.044 (0.4) | -0.017 (0.205) | -0.003 (0.026) |
| **Split S1 80%** | 0.138 (0.155) | 0.418 (0.234) | -0.025 (0.086) | -0.027 (0.081) | 0.376 (0.324) | 0.176 (0.167) | 0.008 (0.021) |
| **AIC** | 145.4 | 166 | 62.3 | 59.2 | 256.1 | 207.7 | -137.4 |

**Table 7A. Expected outcomes by extent of extrapolation and method, evaluated from the different models.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Estimate [confidence interval] | **P(S1=0.6)** | | **P(S1=0.7)** | | **P(S1=0.8)** | |
| Bisection | C+B | Bisection | C+B | Bisection | C+B |
| **Bias, Gauss** | 0.039  [0.557, 1.941] | 0.057  [0.553, 2.026] | 0.043  [0.568, 1.918] | 0.062  [0.552, 2.051] | 0.041  [0.536, 2.025] | 0.059  [0.538, 2.092] |
| **Absolute bias, Gauss (transformed from ln absolute bias scale)** | -0.026  [-0.066, 0.014] | -0.034  [-0.075, 0.007] | -0.038  [-0.078, 0.001] | -0.046  [-0.088,-0.005] | -0.023  [-0.066, 0.020] | -0.031  [-0.074, 0.011] |
| **SDR Gauss** | 3.27  [2.79, 3.85] | 3.19  [2.71, 3.77] | 2.96  [2.52, 3.47] | 2.89  [2.44, 3.41] | 2.82  [2.38, 3.35] | 2.75  [2.32, 3.27] |
| **SDR Gamma** | 3.46  [2.97, 4.02] | 3.36  [2.88, 3.93] | 3.10  [2.67, 3.60] | 3.02  [2.58, 3.53] | 3.02  [2.58, 3.55] | 2.94  [2.50, 3.46] |
| **lnKL Gauss** | 0.08  [-0.21, 0.37] | 0.21  [-0.08, 0.51] | 0.24  [-0.05, 0.52] | 0.37  [0.08, 0.67] | 0.02  [-0.29, 0.32] | 0.15  [-0.16, 0.47] |
| **lnKL Gamma** | 0.12  [-0.31, 0.56] | 0.54  [0.10, 0.98] | 0.43  [-0.01, 0.86] | 0.84  [0.39, 1.29] | 0.07  [-0.39, 0.52] | 0.49  [0.01, 0.96] |
| **KL (transformed from lnKL scale), Gauss** | 1.08  [0.81, 1.44] | 1.24  [0.92, 1.66] | 1.27  [0.95, 1.69] | 1.45  [1.08, 1.96] | 1.02  [0.75, 1.38] | 1.17  [0.85, 1.59] |
| **KL (transformed from lnKL scale), Gamma** | 1.13  [0.73, 1.75] | 1.72  [1.11, 2.68] | 1.53  [0.99, 2.36] | 2.33  [1.48, 3.65] | 1.07  [0.68, 1.69] | 1.62  [1.01, 2.60] |

**Descriptive summaries from experiment 3.1**

Figure 5A shows the distribution of bias, SD ratios and KL divergence in initial priors. Bias is approximately symmetrical and close to zero. Participants were predominantly underconfident – this is consistent with findings from experiments 1 and 2 given that participants played a high precision game.



**Figure 5A. Overview of participants’ initial priors**

**Further analyses and results for experiment 3.1**

**Empirical summaries**

Probability of revision was compared between different types of group feedback and different elicitation methods. The former is presented in section 4.4, while 8A shows the probability of revision for each elicitation method.

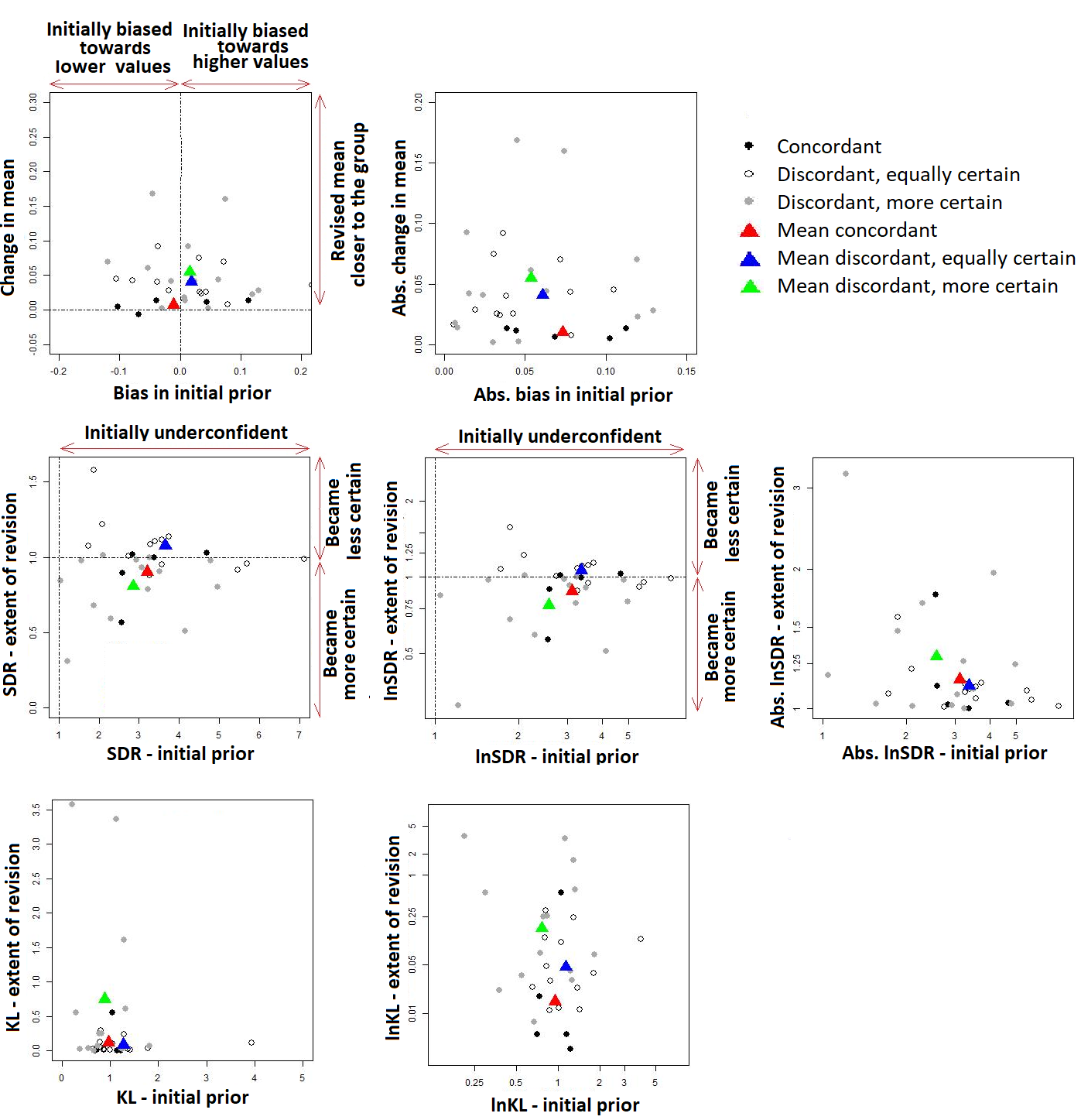
**Table 8A. The probability of revision, per elicitation method.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Concordant (N=25) | | Discordant, equally uncertain (N=24) | | Discordant, more certain (N=23) | |
| Bisection (n=13) | C + B (n=12) | Bisection (n=13) | C + B (n=11) | Bisection (n=11) | C + B (n=12) |
| Proportion who revised their prior (n) | 0% (0) | 41.7% (5) | 53.8% (7) | 54.5% (6) | 45.5% (5) | 75% (9) |

The accuracy of initial priors in those who did and did not revise their priors, are shown in Table 9A. Participants’ extent of revision, for different levels of extrapolation are summarised in 10A. The values in the table summarise the data shown in 6A, and support the findings presented in Chapter 8.

**Table 9A. Accuracy of initial priors compared between participants who did and did not revise their priors: mean (SD) [median and (interquartile range)] of accuracy metric over participants**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Mean (SD)**  **[median (IQR)]** | **Concordant (N=25)** | | **Discordant, equally uncertain (N=24)** | | **Discordant, more certain (N=23)** | |
| **Revised (n=5)** | **Not revised (n=20)** | **Revised (n=13)** | **Not revised (n=11)** | **Revised (n=14)** | **Not revised (n=9)** |
| **Bias** | -0.011 (0.088)  [-0.039 (-0.068 to 0.044)] | -0.004 (0.068)  [0.005 (-0.017 to 0.025)] | 0.018 (0.081)  [0.031 (-0.036 to 0.043)] | 0.013 (0.067)  [0.02 (-0.004 to 0.044)] | 0.016 (0.068)  [0.011 (-0.027 to 0.059)] | 0.033 (0.054)  [0.032 (0.003 to 0.074)] |
| **Absolute bias** | 0.073 (0.033)  [0.068 (0.044 to 0.103)] | 0.044 (0.05)  [0.019 (0.009 to 0.064)] | 0.061 (0.055)  [0.038 (0.033 to 0.078)] | 0.05 (0.045)  [0.036 (0.018 to 0.067)] | 0.054 (0.043)  [0.046 (0.017 to 0.072)] | 0.048 (0.039)  [0.035 (0.014 to 0.074)] |
| **SDR** | 3.21 (0.89)  [2.84 (2.59 to 3.38)] | 3.00 (1.12)  [3.01 (1.95 to 3.89)] | 3.66 (1.58)  [3.40 (2.75 to 3.75)] | 2.80 (1.13)  [2.86 (1.87 to 3.65)] | 2.85 (1.24)  [3.00 (1.93 to 3.44)] | 2.34 (1.03)  [2.57 (1.77 to 3.36)] |
| **lnSDR** | 1.14 (0.25)  [1.04 (0.95 to 1.22)] | 1.03 (0.39)  [1.10 (0.67 to 1.36)] | 1.21 (0.42)  [1.23 (1.01 to 1.32)] | 0.94 (0.46)  [1.05 (0.62 to 1.29)] | 0.95 (0.49)  [1.097 (0.656 to 1.235)] | 0.72 (0.64)  [0.94 (0.57 to 1.21)] |
| **KL** | 0.97 (0.24)  [1.05 (0.73 to 1.15)] | 0.87 (0.61)  [0.84 (0.55 to 1.08)] | 1.282 (0.861)  [1.007 (0.816 to 1.366)] | 0.79 (0.49)  [0.96 (0.34 to 1.03)] | 0.89 (0.46)  [0.81 (0.58 to 1.25)] | 1.648 (2.68)  [0.819 (0.56 to 1.04)] |
| **lnKL** | -0.06 (0.26)  [0.04 (-0.31 to 0.14)] | -0.37 (0.76)  [-0.17 (-0.60 to 0.07)] | 0.12 (0.47)  [0.01 (-0.20 to 0.31)] | -0.53 (0.92)  [-0.05 (-1.23 to 0.03)] | -0.27 (0.63)  [-0.22 (-0.55 to 0.22)] | -0.13 (1.03)  [-0.20 (-0.58 to 0.04)] |



**Figure 6A. Within-participant comparison of accuracy and extent of revision, for different types of group summaries.**

**Table 10A. Aggregate comparison of the extent of revision, for different types of group summaries.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Concordant (N=5) | Discordant, equally uncertain (N=13) | Discordant, more certain (N=14) |
| Mean extent of revision in those who revised (sd)  [median (IQR)] | **Change in mean** | 0.007 (0.008)  [0.012 (0.005 to 0.013)] | 0.041 (0.025)  [0.036 (0.026 to 0.045)] | 0.055 (0.053)  [0.042 (0.019 to 0.068)] |
| **Abs. change in mean** | 0.01 (0.004)  [0.012 (0.006 to 0.013)] | 0.041 (0.025)  [0.036 (0.026 to 0.045)] | 0.055 (0.053)  [0.042 (0.019 to 0.068)] |
| **SDR** | 0.90 (0.20)  [1.00 (0.89 to 1.02)] | 1.08 (0.18)  [1.08 (0.96 to 1.12)] | 0.81 (0.21)  [0.88 (0.71 to 0.98)] |
| **lnSDR** | -0.13 (0.25)  [0.00 (-0.11 to 0.02)] | 0.06 (0.15)  [0.07 (-0.04 to 0.11)] | -0.26 (0.34)  [-0.13 (-0.35 to -0.02)] |
| **Absolute lnSDR** | 0.15 (0.24)  [0.03 (0.02 to 0.11)] | 0.11 (0.12)  [0.09 (0.05 to 0.13)] | 0.29 (0.33)  [0.13 (0.02 to 0.35)] |
| **KL** | 0.12 (0.25)  [0.01 (0.01 to 0.02)] | 0.08 (0.09)  [0.04 (0.02 to 0.12)] | 0.75 (1.23)  [0.16 (0.04 to 0.60)] |
| **lnKL** | -4.20 (2.12)  [-5.29 (-5.30 to -4.04)] | -3.06 (1.15)  [-3.28 (-3.77 to -2.14)] | -1.78 (1.95)  [-1.99 (-3.31 to -0.52)] |

**Modelling**

*Methods*

The probability of revision was modelled as follows:

|  |  |
| --- | --- |
| lnOddsj= b0 + b1\*y0j + b2\*I(Sj==2) + b3\*I(Sj==3) + b4\*Xj + ej, | (Equation 3) |
|  |  |

j is the subject, where lnOdds is the log of the odds of revision, y0j is individual j’s accuracy for the initial distribution (bias, ln SDR or ln KL), Sj is the group individual j was randomised to, I() the indicator function that assumes a value of 1 of the enclosed statement is true and 0 if false, and Xj the method of elicitation used by subject j (0 if chips and bins is used or 1 if bisection is used, here a blocking factor). In experiment 3.2 there were only two randomisation groups hence the term with b3 is redundant.

Four different models were compared, three of these used a single independent term to express accuracy in y0j, each using each of the different metrics (lnKL divergence, bias and lnSDR). The forth model used two independent terms for bias and uncertainty.

*Results*

The logistic regression model coefficients are shown in 11A. The coefficients represent the change in the log of the odds of revision, where positive values suggest that the coefficient increases the odds of revision and vice versa. Four different models were compared, where each used different metrics of probabilistic accuracy. Consistently, the models suggest that both the type of group summary and the elicitation method affected participants’ probability of revision. Participants were more likely to revise their priors when the group was discordant, when the group was more certain, and when using Chips and Bins compared to Bisection. The model coefficients suggest that participants’ level of uncertainty (lnSDR) has an effect on the likelihood of revision, with more uncertain participants showing a higher likelihood of revision. The effect of bias and KL divergence was not shown to be different from zero (wide standard errors in relation to the mean).

The fitted models also enable us to compare estimates and confidence intervals for expected outcomes between groups, in 12A.

**Table 11A. Logistic regression model coefficients, showing the effect of the type of group summary, initial accuracy and the elicitation method on the log of the odds of revision. Each model uses a different measure of accuracy.**

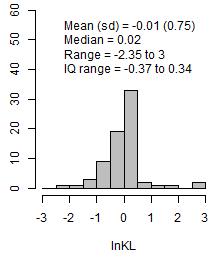
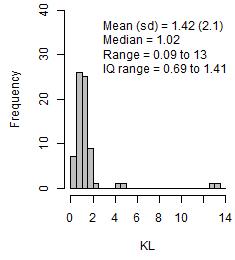
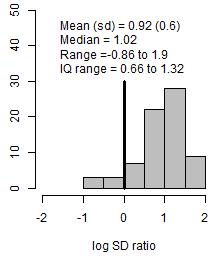
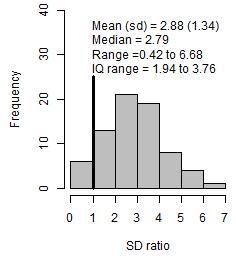
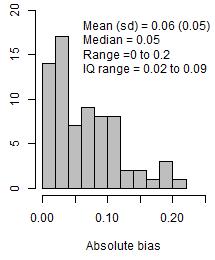
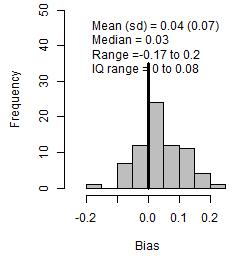
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 | Model 4 |
|  | Log Odds ratio (SE) shown for coefficients and intercept | | | |
| Intercept | -1.943 (0.627) | -2.38 (0.726) | -4.072 (1.125) | -4.512 (1.216) |
| Discordant, equally uncertain | 1.676 (0.689)\* | 1.707 (0.687)\* | 1.802 (0.71)\* | 1.801 (0.72)\* |
| Discordant, more certain | 1.953 (0.705)\* | 1.949 (0.6995)\* | 2.4 (0.768)\* | 2.463 (0.786)\* |
| Initial accuracy, lnKL | 0.678 (0.414) | - | - | - |
| Initial accuracy, absolute bias | - | 6.294 (6.148) | - | 7.396 (6.445) |
| Initial accuracy, lnSDR | - | - | 1.592 (0.675)\* | 1.559 (0.612)\* |
| Chips and Bins | 1.282 (0.561)\* | 1.135 (0.542)\* | 1.619 (0.613)\* | 1.638 (0.681)\* |
| AIC | 90.9 | 92.9 | 87.5 | 88.1 |

**Table 12A. Expected probability of revision, per type of group feedback and elicitation method.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Estimate [confidence interval] | Concordant | | Discordant, equally uncertain | | Discordant, more certain | |
| Bisection | C+B | Bisection | C+B | Bisection | C+B |
| Model 1 | 0.109  [0.034, 0.297] | 0.306  [0.129, 0.567] | 0.395  [0.198, 0.633] | 0.702  [0.447, 0.872] | 0.463  [0.237, 0.705] | 0.756  [0.509, 0.903] |
| Model 2 | 0.114  [0.036, 0.306] | 0.286  [0.120, 0.542] | 0.415  [0.214, 0.648] | 0.688  [0.437, 0.862] | 0.475  [0.249, 0.712] | 0.738  [0.497, 0.889] |
| Model 3 | 0.077  [0.021, 0.247] | 0.298  [0.124, 0.559] | 0.337  [0.154, 0.588] | 0.720  [0.463, 0.884] | 0.481  [0.243, 0.728] | 0.824  [0.579, 0.941] |
| Model 4 | 0.077  [0.021, 0.249] | 0.284  [0.114, 0.550] | 0.336  [0.152, 0.587] | 0.706  [0.444, 0.879] | 0.495  [0.249, 0.743] | 0.823  [0.578, 0.941] |

**Descriptive summaries from experiment 3.2**

Figure 7A shows the distribution of bias, SD ratios and KL divergence in initial priors. Bias is approximately symmetrical and close to zero. Participants were predominantly under confident – this is consistent with findings from experiments 1, 2, and 3.1, given that participants played a high precision game.



**Figure 7A. Overview of initial priors**

**Further analyses and results for experiment 3.2**

Experiment 3.2 explored how individuals reviewed their own probabilistic assessments when presented with Delphi-type summaries. Individuals’ revision of probabilistic assessments was measured in terms of the probability of revision, and the extent of revision in those who revised. This appendix provides details modelling of the probability of revision.

**Methods**

As described for experiment 3.1

*Results*

The logistic regression model coefficients are shown in 13A. Consistently, the models suggest that participants are more likely to revise their priors when individuals in the group were consistent with each other, and when using Chips and Bins compared to Bisection (narrow standard error in relation to the mean).

The fitted models also enable us to compare estimates and confidence intervals for expected outcomes between groups, in 14A.

**Table 13A. Logistic regression model coefficients, showing the effect of the type of group summary, initial accuracy and the elicitation method on the log of the odds of revision.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 | Model 4 |
| Log Odds ratio (SE) shown for coefficients and intercept | | | |
| Intercept | -0.659 (0.454) | -0.897 (0.575) | -1.331 (0.715) | -1.562 (0.822) |
| Inconsistent | -1.657 (0.605) | -1.629 (0.606) | -1.512 (0.581) | -1.612 (0.61) |
| Initial accuracy, lnKL | 0.297 (0.357) | - | - | - |
| Initial accuracy, Absolute bias | - | 3.622 (5.904) | - | 3.722 (5.976) |
| Initial accuracy, lnSDR | - | - | 0.619 (0.5) | 0.628 (0.507) |
| Chips and Bins | 1.814 (0.589) | 1.821 (0.591) | 1.903 (0.598) | 1.96 (0.612) |
| AIC | 85.3 | 85.6 | 84.3 | 85.9 |

**Table 14A. Expected probability of revision, per type of feedback and elicitation method.**

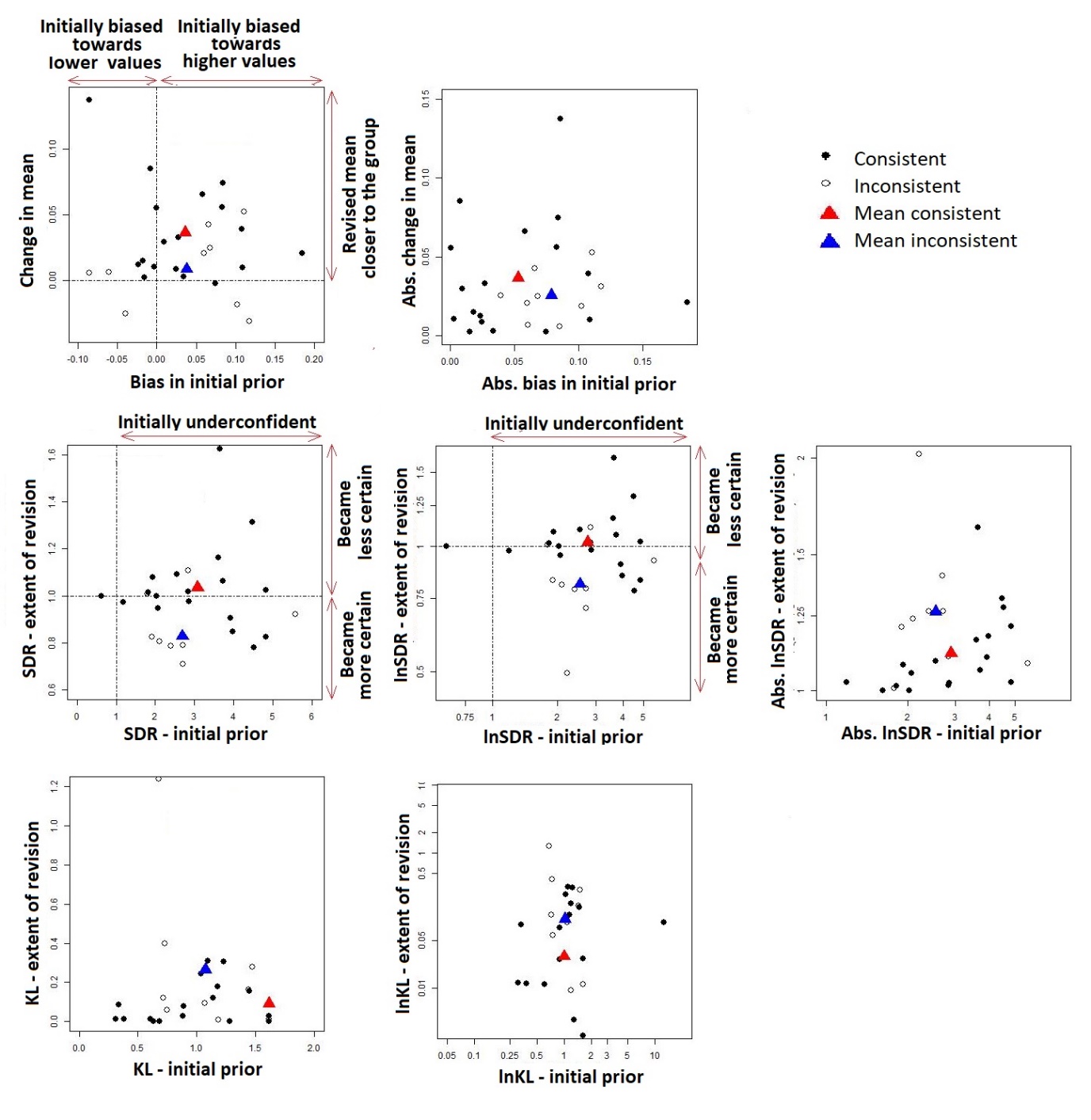
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Estimate [confidence interval] | Consistent | | Inconsistent | |
| Bisection | C+B | Bisection | C+B |
| Model 1 | 0.340 [0.175, 0.557] | 0.760 [0.539, 0.895] | 0.090 [0.029, 0.245] | 0.376 [0.200, 0.592] |
| Model 2 | 0.340 [0.175, 0.556] | 0.761 [0.535, 0.898] | 0.092 [0.030, 0.247] | 0.384 [0.207, 0.598] |
| Model 3 | 0.318 [0.157, 0.539] | 0.758 [0.536, 0.894] | 0.093 [0.032, 0.242] | 0.408 [0.222, 0.625] |
| Model 4 | 0.321 [0.158, 0.544] | 0.771 [0.545, 0.904] | 0.086 [0.028, 0.237] | 0.401 [0.216, 0.619] |

Table 15A shows the accuracy of initial priors in those who did and did not revise. This analysis shows no evidence that probabilistic accuracy is associated with participants’ likelihood of revision.

**Table 15A. Outcomes in participants who did and did not revise their priors.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mean (sd)**  **[median (IQR)]** | **Consistent (N=34)** | | **Inconsistent (N=38)** | |
| **Revised (n=18)** | **Not revised (n=16)** | **Revised (n=9)** | **Not revised (n=29)** |
| **Bias** | 0.036 (0.064)  [0.026 (-0.007 to 0.081)] | 0.023 (0.07)  [0.017 (-0.005 to 0.048)] | 0.038 (0.078)  [0.066 (-0.04 to 0.102)] | 0.046 (0.082)  [0.039 (0.006 to 0.104)] |
| **Absolute bias** | 0.053 (0.049)  [0.03 (0.016 to 0.084)] | 0.052 (0.051)  [0.033 (0.012 to 0.077)] | 0.079 (0.026)  [0.068 (0.061 to 0.102)] | 0.074 (0.057)  [0.049 (0.034 to 0.106)] |
| **SDR** | 3.08 (1.28)  [3.23 (2.04 to 3.97)] | 2.73 (1.55)  [2.51 (1.71 to 3.37)] | 2.69 (1.14)  [2.40 (2.10 to 2.71)] | 2.89 (1.36)  [3.03 (1.95 to 3.64)] |
| **lnSDR** | 1.01 (0.55)  [1.17 (0.71 to 1.38)] | 0.82 (0.68)  [0.92 (0.53 to 1.21)] | 0.93 (0.34)  [0.88 (0.74 to 1.00)] | 0.90 (0.66)  [1.11 (0.67 to 1.29)] |
| **Absolute lnSDR** | 1.07 (0.42)  [1.17 (0.71 to 1.38)] | 0.94 (0.49)  [0.92 (0.73 to 1.21)] | 0.93 (0.34)  [0.88 (0.74 to 1.00)] | 1.04 (0.41)  [1.11 (0.83 to 1.29)] |
| **KL** | 1.61 (2.79)  [1.07 (0.65 to 1.27)] | 0.87 (0.46)  [0.79 (0.66 to 1.13)] | 1.07 (0.37)  [1.07 (0.73 to 1.44)] | 1.72 (2.45)  [1.09 (0.85 to 1.46)] |
| **lnKL** | -0.01 (0.82)  [0.06 (-0.44 to 0.24)] | -0.35 (0.79)  [-0.24 (-0.43 to 0.12)] | 0.01 (0.35)  [0.07 (-0.32 to 0.36)] | 0.16 (0.75)  [0.09 (-0.16 to 0.38)] |

Participants’ extent of revision, for different levels of extrapolation are summarised in 16A. The values in the table summarise the data shown in 8A, and support the findings presented in Chapter 8.



**Figure 8A. Within-participant comparison of accuracy and extent of revision, for different types of group summaries.**

**Table 16A. Participants’ extent of revision, for different levels of extrapolation**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Consistent (N=18) | Inconsistent (N=9) |
| Mean extent of revision in those who revised (sd)  [median (IQR)] | **Change in mean** | 0.036 (0.037)  [0.025 (0.01 to 0.056)] | 0.009 (0.03)  [0.006 (-0.018 to 0.025)] |
| **Abs. change in mean** | 0.037 (0.036)  [0.025 (0.01 to 0.056)] | 0.025 (0.015)  [0.025 (0.018 to 0.031)] |
| **SDR** | 1.037 (0.192)  [1.008 (0.955 to 1.076)] | 0.828 (0.176)  [0.808 (0.789 to 0.922)] |
| **lnSDR** | 0.022 (0.169)  [0.008 (-0.046 to 0.073)] | -0.211 (0.231)  [-0.214 (-0.238 to -0.082)] |
| **Absolute lnSDR** | 0.111 (0.126)  [0.07 (0.023 to 0.16)] | 0.235 (0.202)  [0.214 (0.103 to 0.238)] |
| **KL** | 0.093 (0.106)  [0.053 (0.012 to 0.148)] | 0.264 (0.386)  [0.12 (0.061 to 0.279)] |
| **lnKL** | -3.527 (1.991)  [-3.069 (-4.463 to -1.919)] | -2.248 (1.59)  [-2.118 (-2.802 to -1.275)] |