Supplemental materials for *Base rate neglect and conservatism in probabilistic reasoning: Insights from eliciting full distributions*

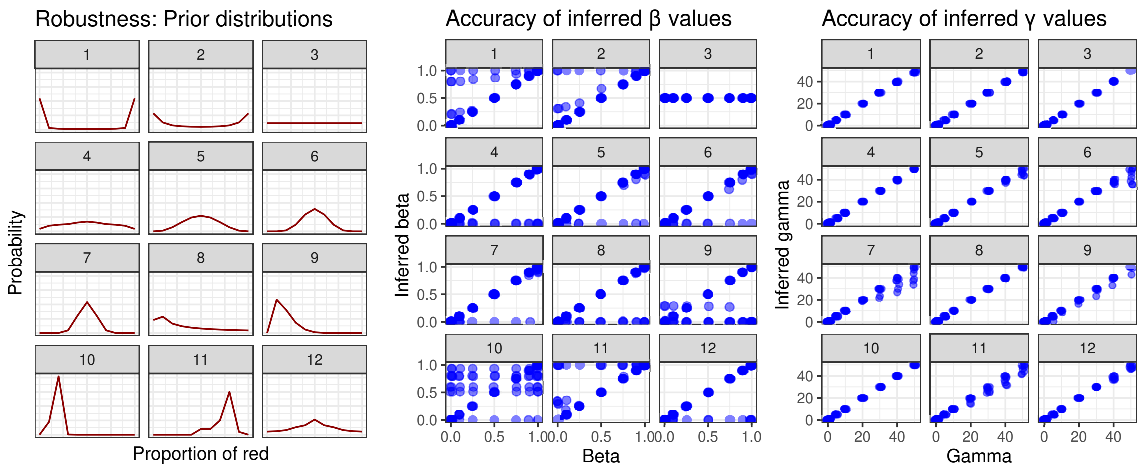
Piers Douglas Lionel Howe[1], Andrew Perfors[1], Bradley Walker[2],
Yoshihisa Kashima[1], and Nicolas Fay[2]
[1]School of Psychological Sciences, University of Melbourne
[2]School of Psychological Science, University of Western Australia

Robustness analysis

In order to evaluate how good this method is in general, we considered a range of priors of different shapes. Shown on the left panel of Figure 1, this included bimodal priors (numbers 1 and 2); a uniform prior (number 3); priors of different strength centred around 50% (numbers 4 through 7); unimodal of different strength centred on the opposite side that the data would be one (numbers 8 through 10); the given prior in Experiment 2, which is unimodal and centred on the side the data would be on (number 11); and the given prior in Experiment 3 (number 12). For each of these priors, we systematically generated posteriors (assuming 4 red and 1 blue chip) for every possible combination of $\beta = [0.001, 0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.99, 0.999]$ and $\gamma = [0.01, 0.1, 0.5, 1, 5, 10, 20, 30, 40, 50]$. We then used our model to infer values of β and γ for each. Figure 1 shows the accuracy of the inference. It is evident that for all priors, inference is very accurate for γ and for all priors except for number 3 (the uniform one), inference is reasonably accurate for β.

Figure 1

Robustness analysis: Accuracy



Note. Accuracy with which the model infers the actual parameters used to generate simulated data, for each of the priors in the left panel and all possible combinations of β and γ. The middle panel compares the inferred β values (y axis) to the actual β values used to generate them, while the right panel compares the inferred γ values (y axis) to the actual γ values used to generate them. For both parameters, the model accurately infers them the majority of the time, as is evident by the fact that most of the parameters are found along the diagonal (with the exception of β for the uniform prior 3, which makes sense because β is the weight between it and a uniform prior and it is meaningless when the prior is uniform.)

Although performance is generally good, it is not perfect, particularly for some β values. In order to understand under what circumstances (i.e., for what combinations of prior, β, and γ values) the model fails to recover the correct parameters, in Figure 2 we plot the points in the parameter space for which the error in each parameter was 0.1 or less.

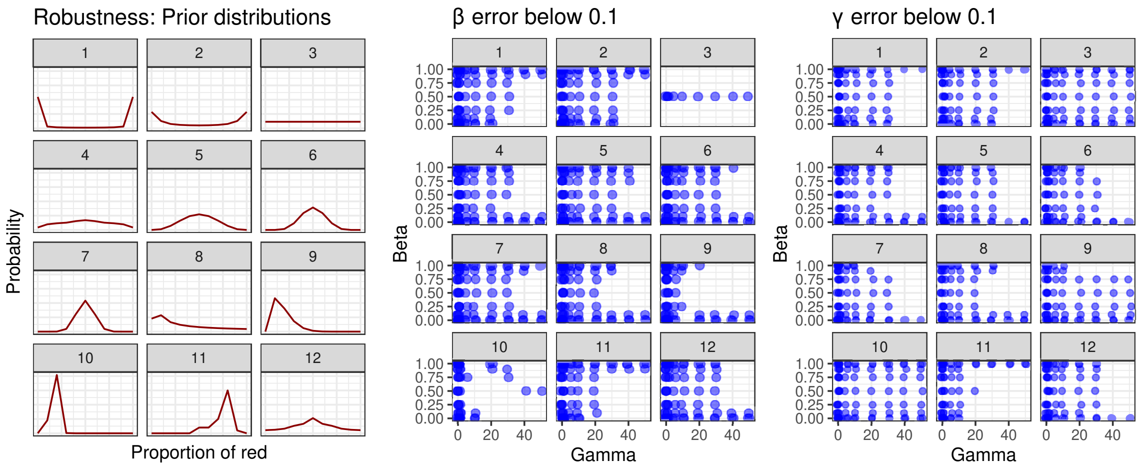
The middle panel shows all of the simulated runs in which the error in the β values was 0.1 or less (i.e., the inferred β value was within 0.1 of the correct β value used to generate the data). Thus, regions of the space with corresponding blue dots indicate the regions in which parameter recovery of β is successful for that prior. It is evident that for all priors except for number 3 (the uniform one), parameter recovery of β is highly accuracy unless γ is large. This makes sense because a large γ corresponds to a substantial overweighting of the likelihood, which means that the influence of any prior gets washed out in the sheer amount of data.

The right panel shows all of the simulated runs in which the error in the γ values was 0.1 or less (i.e., the inferred γ value was within 0.1 of the correct γ value used to generate the data). Thus, regions of the space with corresponding blue dots indicate the regions in which parameter recovery is successful for that prior. Parameter recovery for γ was generally highly successful, although again lowest when γ was high; in this situation the model not only tended to infer $\beta = 1$ no matter what, but found it harder to accurately estimate γ, presumably because when it is that large, the data is overweighted so much that it is impossible to detect small differences in γ.

Importantly for our purposes, very few of our participants overweighted the data that much. Even among those for whom the model inferred $\gamma > 1$, most of those obtained γ values of 10 or less.

Figure 2

Robustness: High-accuracy parameter space



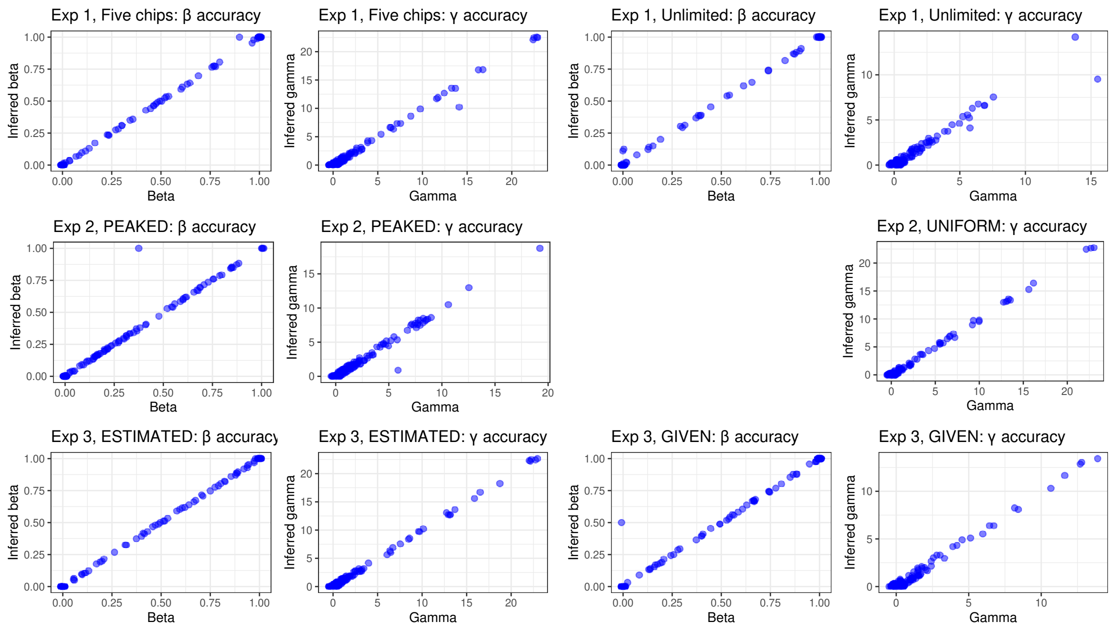
Note. Parameter space for which the model accurately inferred the true parameters used to generate simulated data, for each of the priors in the left panel and all possible combinations of β and γ. The middle panel depicts all of the simulated runs for which the error in the β values was 0.1 or less (thus, blue dots indicate successful parameter recovery of β for the corresponding values of β (*x* axis) and γ (*y* axis)). The right panel depicts all of the simulated runs for which the error in the γ values was 0.1 or less (thus, blue dots indicate successful parameter recovery of γ for the corresponding values). For both parameters, the model was highly accurate at inferring the parameters except when γ was very large, which makes sense because that is when the likelihood is overweighted so much that any effect of the prior is effectively impossible to discern.

Plausibility

The robustness analysis above explores how effectively this model recovers the parameters *in general*, but it can also be useful to determine to what extent it may have plausibly recovered the particular parameters for our participants in each experiment. This is impossible to determine with certainty since we do not know the true parameters, but we can roughly estimate the plausibility of parameter inference. We do this by considering each participant in each of the experiments for whom we estimated β and γ priors (i.e., who reported both priors and posteriors). For each person, we produce a generated posterior based on their prior and their β and γ parameters reported in the main paper. We then present our model with the generated posterior and their prior, and ask it to infer both β and γ. If the model is plausible and the parameters are recoverable, the β and γ we infer based on the simulated posterior should match the parameters we used to simulate it. The results, shown below in Figure 3, show that for all experiments, parameter recovery was excellent. All correlations between the given parameters and the inferred ones were 0.97 or higher.

Figure 3

Plausibility: Parameter recovery for each participant



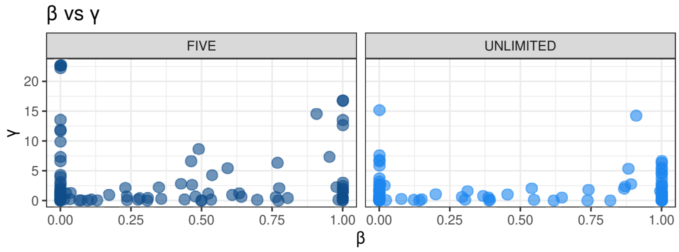
Note. Comparison of the parameter values inferred by the model (y axis) to the actual values used to generate simulated posteriors. Each participant corresponds to one dot, for whom the simulated posteriors were generated based on their prior and the β and γ values the model inferred in the main analysis. For all experiments, parameter inference was highly accurate: correlations between the true parameters and the inferred ones were always 0.97 or higher (all $ps < .0001$).

Experiment 1

Main analyses

Figure 4

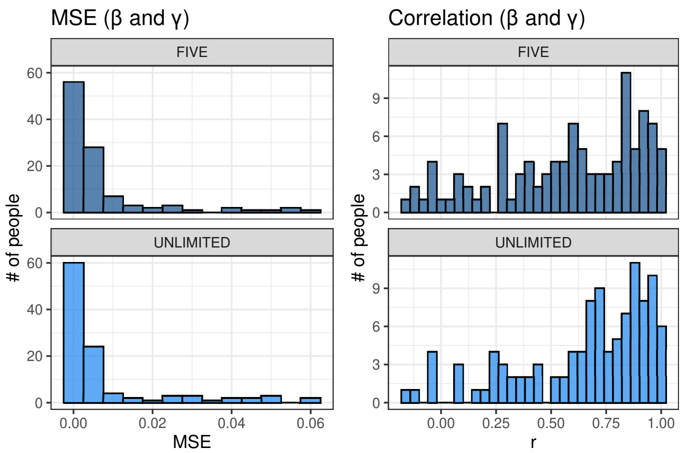
Experiment 1: Relationship between β and γ of individuals after both five chips and an unlimited number of chips



Note. Relationship between fitted β and γ values in the MAIN condition of Experiment 1. Each dot indicates one participant. The plots show that after both five and unlimited chips, there is no systematic relationship between β and γ.

Figure 5

Experiment 1: Goodness of model fit after five chips and after an unlimited number of chips in the MAIN *condition*



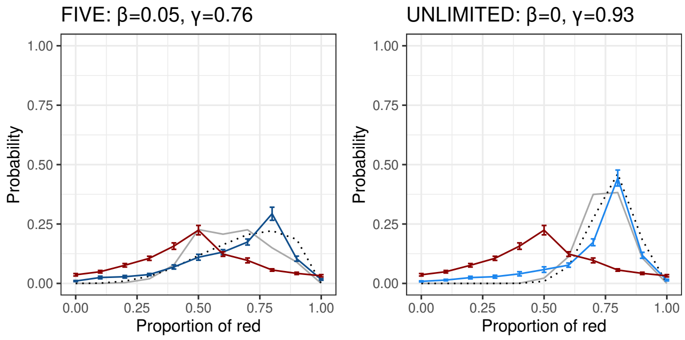
Note. We quantify how well our model fits individuals in two ways. The panel on the left shows the distribution of mean squared error (MSE) between the fitted posteriors (i.e. the posteriors generated by the model) and the participants' reported posteriors. It is evident that most fits were good, with MSE values below 0.01. The panel on the right shows the correlation between the fitted posteriors and the participants' reported posteriors. Again, most correlations (r) were strong, indicating that most participants were fit well.

Good participants

Below we report all analyses on the subset of participants (76 of the original 107) who were fit well by the model, i.e., who had an MSE less than 0.01. We'll call these the GOOD ones.

Figure 6

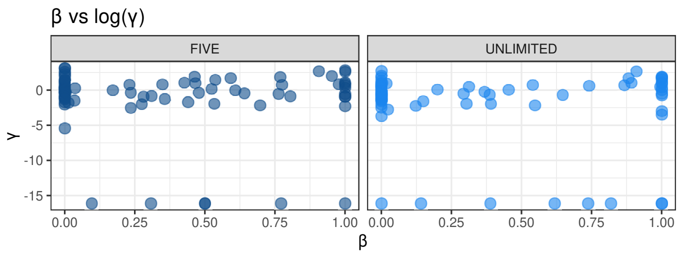
Experiment 1 GOOD *participants: Aggregate best-fit estimates*



Note. Reported distributions for prior beliefs (red line) as well as after seeing five (dark blue line, left panel) and unlimited (light blue line, right panel) chips. In both panels, the grey line indicates the optimal Bayesian prediction given the prior, while the black dotted line indicates the prediction of the line of best fit based on the inferred parameters β and γ. In both panels, as in the full dataset, β is around zero, indicating that the aggregate posterior was best fit assuming that participants disregarded their reported prior (i.e., showed almost complete base rate neglect). The value for γ indicates some conservatism on average in both conditions, somewhat less than was observed in the full dataset.

Figure 7

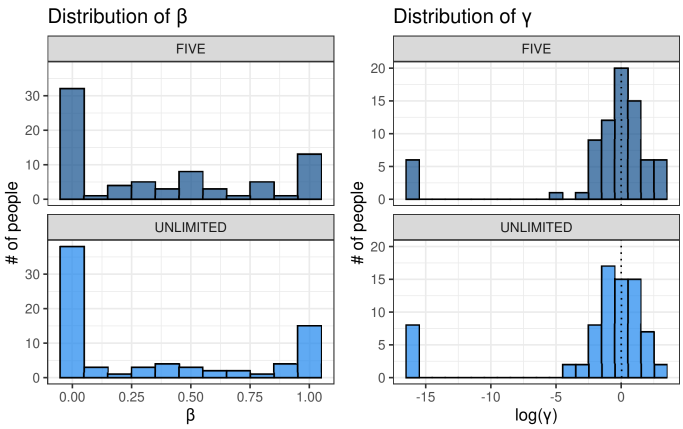
Experiment 1 GOOD *participants: Relationship between* β *and* γ



Note. Relationship between obtained β and γ values. Each dot indicates one participant. The plots show that after both five and unlimited chips, there is no systematic relationship between β and γ.

Figure 8

Experiment 1 GOOD *participants: Distribution of fitted* β *and* γ *values across individuals*



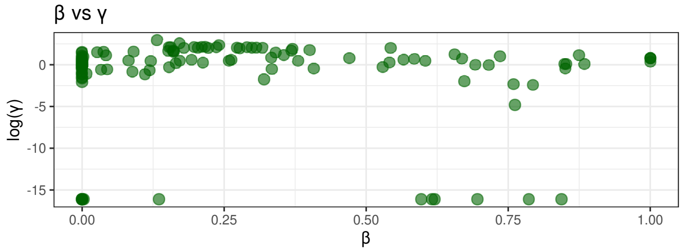
Note. Histograms showing the distribution of best-fit β and γ values across individuals. The panels on the left show that the majority of people showed a moderate or large amount of base rate neglect; their inferences were best described with β values less than one and often close to zero. 43.4% of people after seeing five chips and 50% after seeing an unlimited number of additional chips had $\beta < 0.1$, and 18.4% after seeing five chips and 21.1% after seeing an unlimited number had $\beta > 0.9$. The value for γ indicates a range of conservatism, with 52.6% after seeing five chips and 57.9% after seeing an unlimited number with $\gamma < 1$.

Experiment 2

Main analyses

Figure 9

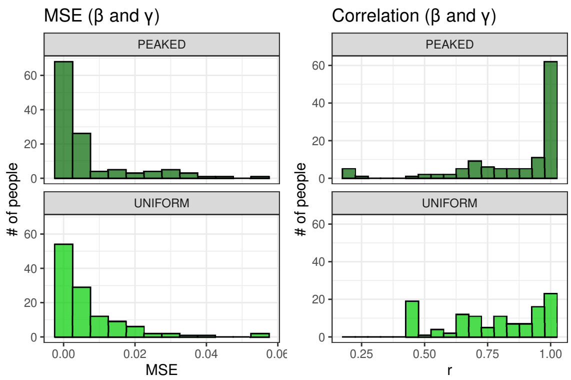
Experiment 2: Relationship between β and γ



Note. Relationship between β and γ values for the PEAKED condition. Each dot indicates one participant. There is no systematic relationship between β and γ, but the distribution of the two parameters is strikingly different than in the other conditions. There were very few participants with extremely high γ values (above 15 or so), unlike in the other experiments. In those experiments, a peaked posterior at 80% could only be accounted for with this very high γ value, since β changed the shape of the distribution. In Experiment 2, however, a peaked posterior at 80% could also be accounted for with a somewhat more moderate γ value (above 1 but less than 15) in conjunction with a non-zero β. This is one indication of the non-identifiability problem discussed in the main paper. (The UNIFORM condition is not shown because β is not inferred there.)

Figure 10

Experiment 2: Goodness of model fit



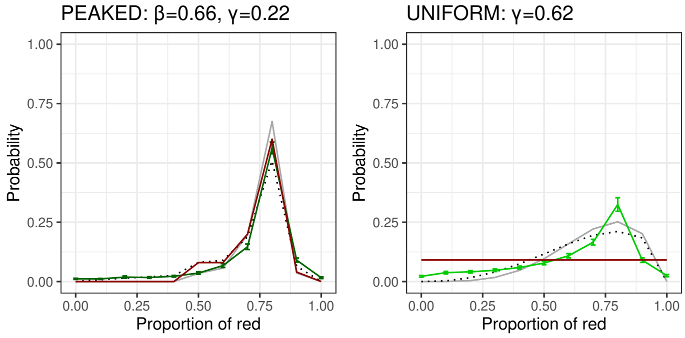
Note. The panel on the left shows the distribution of mean squared error (MSE) between the fitted posteriors and the reported posteriors. The panel on the right shows the correlation between the fitted and reported posteriors. Fits in both conditions were good but noticeably better fit in the PEAKED condition, presumably because it had two free parameters instead of one.

Good participants

Below we report all analyses on the subset of participants (189 of the original 239) who were fit well by the model, i.e., who had an MSE less than 0.01. As before, we'll call these the GOOD ones.

Figure 11

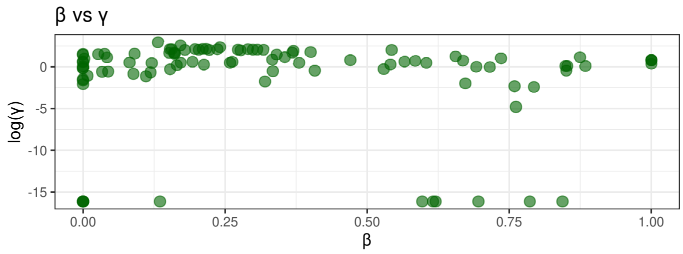
Experiment 2 GOOD participants: Aggregate best-fit estimates



Note. Reported distributions in the PEAKED (dark green line, left panel) and UNIFORM (light green line, right panel) conditions. The grey line indicates the optimal Bayesian prediction given the prior, while the black dotted line indicates the prediction of the line of best fit based on the inferred parameters β and γ. In the PEAKED condition, β indicates a moderate degree of base rate neglect: higher than in Experiment 1 but still present. The value for γ again indicates moderate conservatism, although given the identifiability issues these parameters should be interpreted with caution.

Figure 12

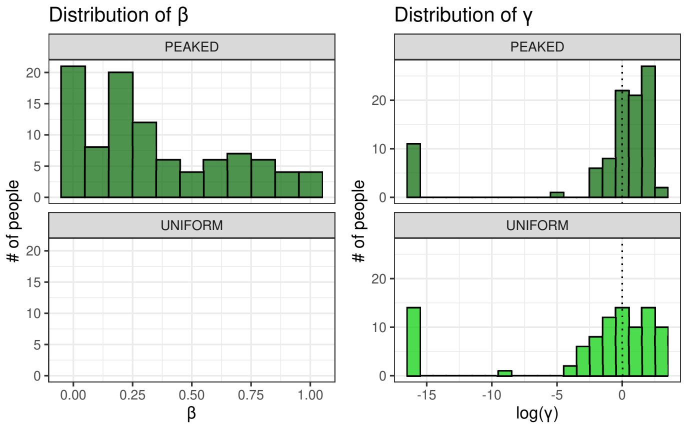
Experiment 2 GOOD participants: Relationship between β and γ



Note. Relationship between obtained β and γ values. Each dot indicates one participant. As in the full dataset, there was no systematic relationship between β and γ.

Figure 13

Experiment 2 GOOD *participants: Distribution of fitted* β *and* γ *values across individuals*



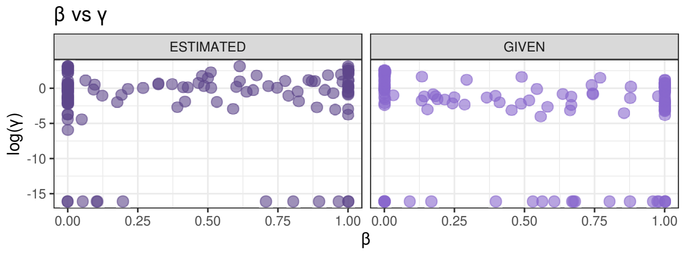
Note. Histograms showing the distribution of best-fit β and γ values across individuals in the PEAKED and UNIFORM conditions. The panels on the left show that the majority of people showed some amount of base rate neglect, but less *extreme* base rate neglect in this sample than in the full dataset (24.5% of people had $\beta < 0.1$, compared to 38.8% in the full dataset); however, only a small minority showed none at all (4.1% had $\beta > 0.9$, compared to 3.3% in the full dataset). The value for γ indicates a range of conservatism in both conditions (though less in PEAKED), with 34.7% in PEAKED and 54.9% in UNIFORM with $\gamma < 1$. This probably results, at least in part, from the unidentifiability issues discussed elsewhere.

Experiment 3

Main analyses

Figure 14

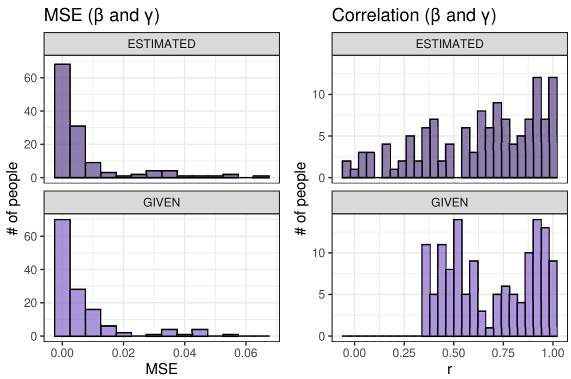
Experiment 3: Relationship between β and γ



Note. Relationship between obtained β and γ values. Each dot indicates one participant. The plots show that in both ESTIMATED and GIVEN conditions, there is no systematic relationship between β and γ.

Figure 15

Experiment 3: Goodness of fit



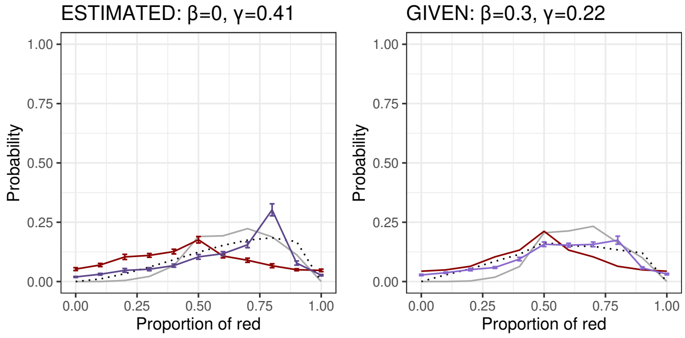
Note. We quantify how well our model fits individuals in two ways. The panel on the left shows the distribution of mean squared error (MSE) between the fitted posteriors and the reported posteriors. It is evident that most fits were good, with MSE values below 0.01. The panel on the right shows the correlation between the fitted posteriors and the reported posteriors. Again, most correlations (*r*) were strong, indicating that most participants were fit well.

Good participants

Below we report all analyses on the subset of participants (213 of the original 261) who were fit well by the model, i.e., who had an MSE less than 0.01. As before, we'll call these the GOOD ones.

Figure 16

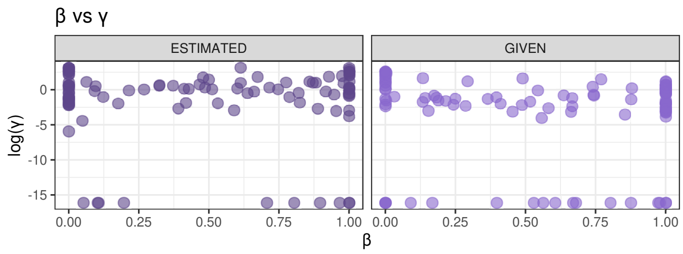
Experiment 3 GOOD participants: Aggregate best-fit estimates



Note. Reported distributions for prior beliefs (red line) and posteriors in the ESTIMATED (dark purple line, left panel) and GIVEN (light purple line, right panel) conditions. In both panels, the grey line indicates the optimal Bayesian prediction given the prior, while the black dotted line indicates the prediction of the line of best fit based on the inferred parameters β and γ. As in the full dataset, β is around zero in the ESTIMATED condition and slightly higher in the GIVEN condition. The value for γ indicates a moderate degree of conservatism on average in both conditions.

Figure 17

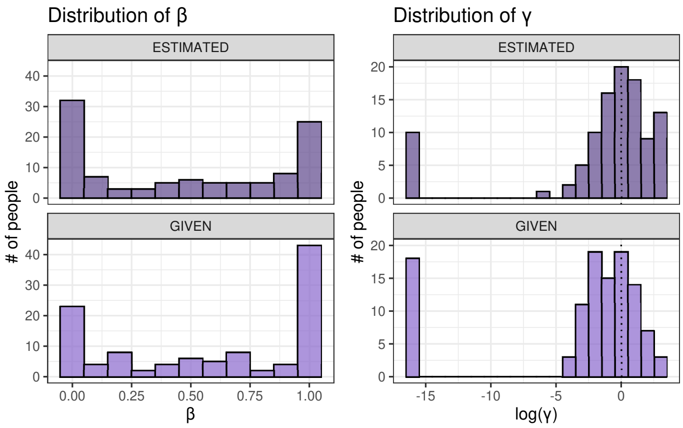
Experiment 3 GOOD participants: Relationship between β and γ



Note. Relationship between obtained β and γ values. Each dot indicates one participant. The plots show that in both ESTIMATED and GIVEN conditions, there is no systematic relationship between β and γ.

Figure 18

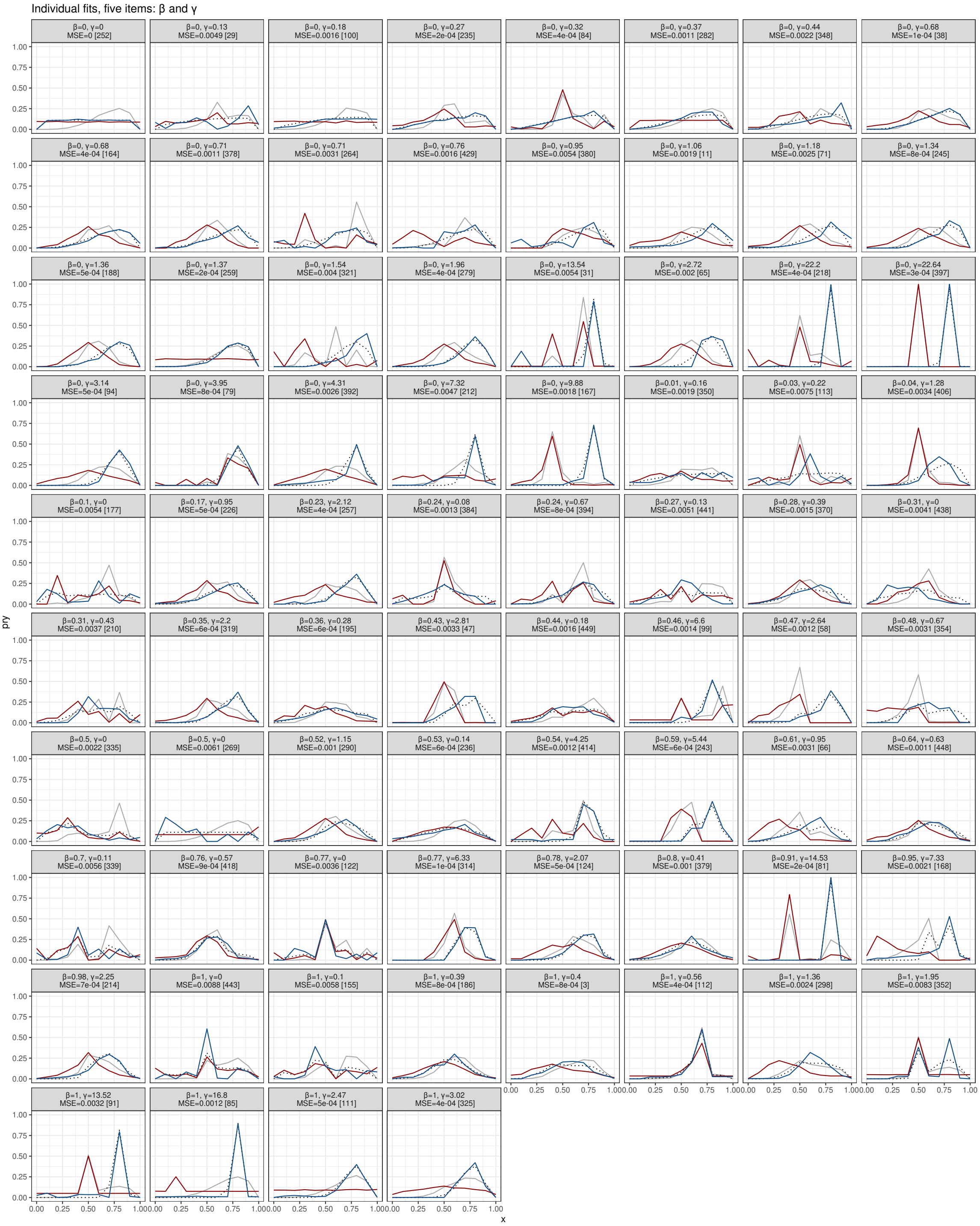
Experiment 3 GOOD *participants: Distribution of fitted* β *and* γ *values across individuals*



Note. Histograms showing the distribution of best-fit β and γ values across individuals. As in the full dataset, many people showed a moderate or large amount of base rate neglect in the ESTIMATED condition, in the GIVEN condition many people showed no neglect at all. 34.6% of people in ESTIMATED and 22% in GIVEN had $\beta < 0.1$, and 26.9% in ESTIMATED and 39.4% in GIVEN had $\beta > 0.9$. The value for γ indicates a range of conservatism, with 48.1% in ESTIMATED and 68.8% in GIVEN with $\gamma < 1$.

Figure 19

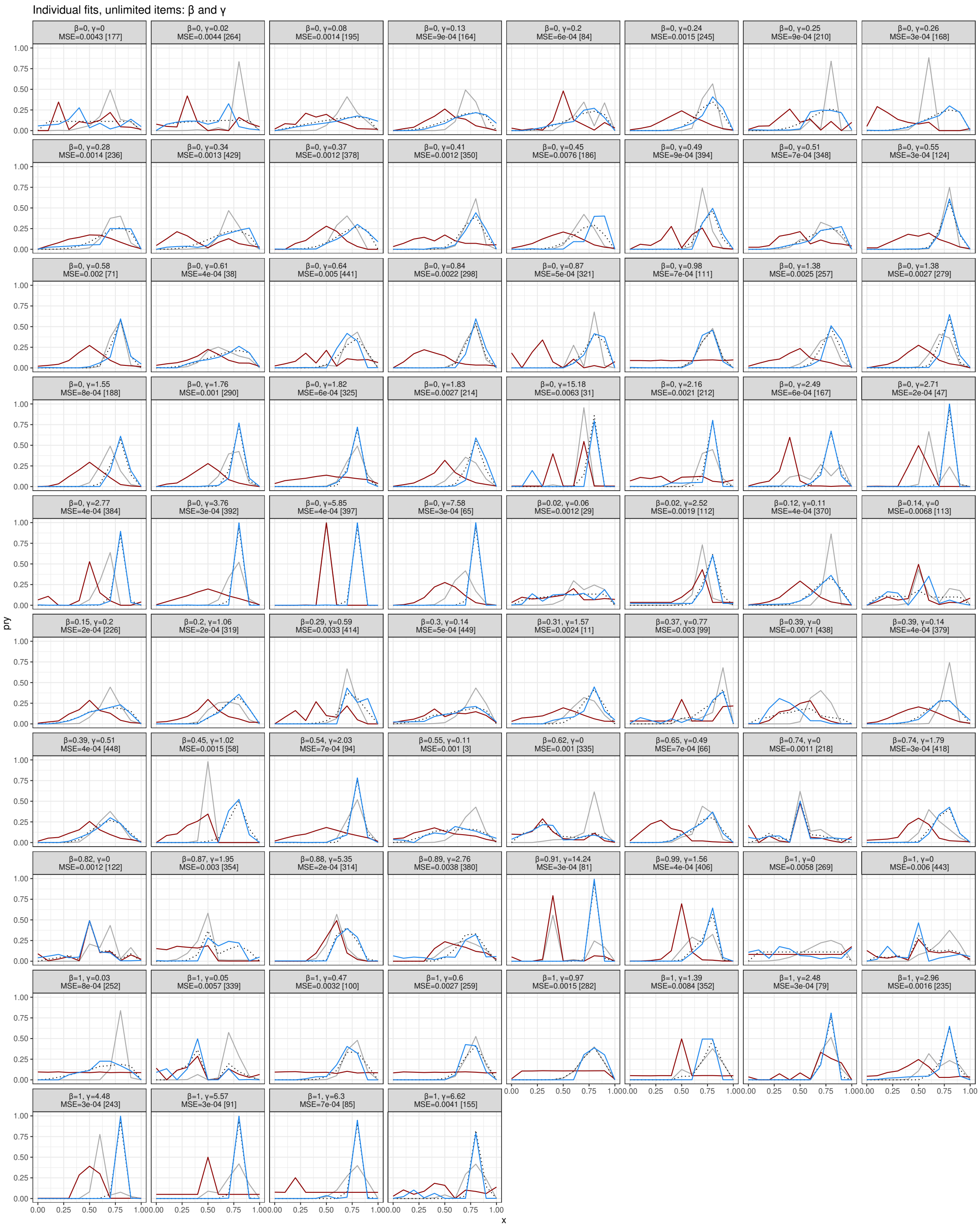
Experiment 1 GOOD *participants: Each person after five chips, fitting* β *and* γ



Note. Each good individual participant after receiving five chips. The red line is the reported prior, the dark blue line is reported posterior, the grey line is the posterior obtained by a well-calibrated Bayesian reasoner with that prior and $\beta = \gamma = 1$, and the dotted black line is the posterior obtained by the best-fit values of β and γ for that person. The grey label for each panel contains those values as well as the mean squared error of the fit (MSE, with 0 being perfect). The number in square brackets is the participant ID.

Figure 20

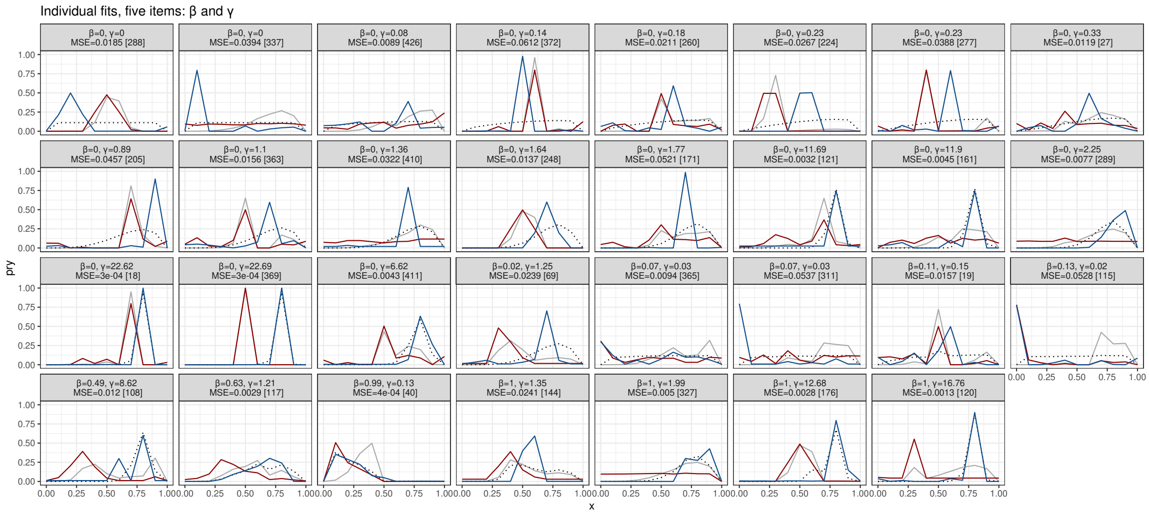
Experiment 1 GOOD *participants: Each person after unlimited chips, fitting* β *and* γ



Note. Each good individual participant after receiving unlimited chips. The red line is the reported prior, the light blue line is reported posterior, the grey line is the posterior obtained by a well-calibrated Bayesian reasoner with that prior and $\beta = \gamma = 1$, and the dotted black line is the posterior obtained by the best-fit values of β and γ for that person. The grey label for each panel contains those values as well as the mean squared error of the fit (MSE, with 0 being perfect). The number in square brackets is the participant ID.

Figure 21

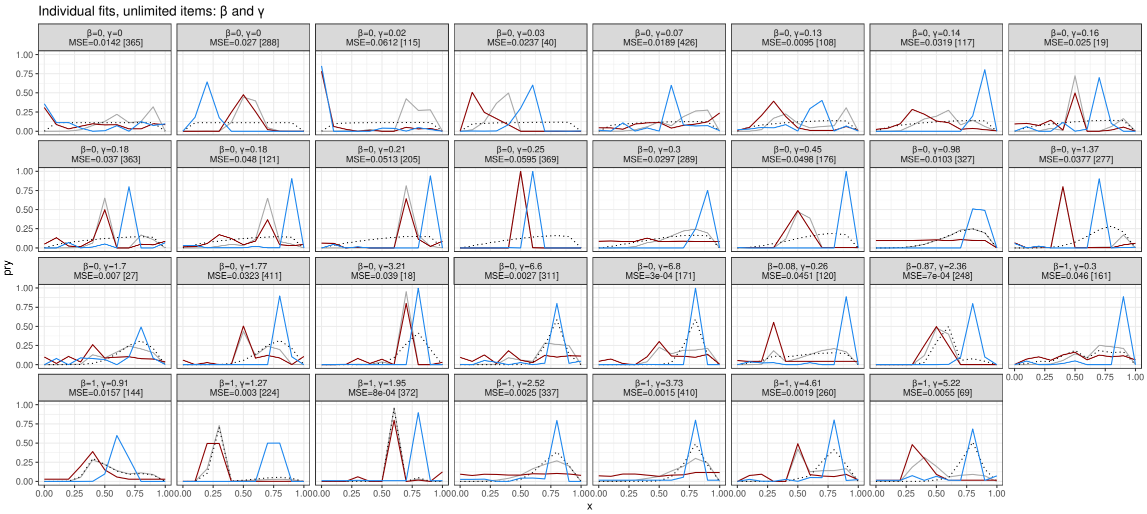
Experiment 1 BAD *participants: Each person after five chips, fitting* β *and* γ



Note. Each bad individual participant after receiving five chips. The red line is the reported prior, the dark blue line is reported posterior, the grey line is the posterior obtained by a well-calibrated Bayesian reasoner with that prior and $\beta = \gamma = 1$, and the dotted black line is the posterior obtained by the best-fit values of β and γ for that person. The grey label for each panel contains those values as well as the mean squared error of the fit (MSE, with 0 being perfect). The number in square brackets is the participant ID.

Figure 22

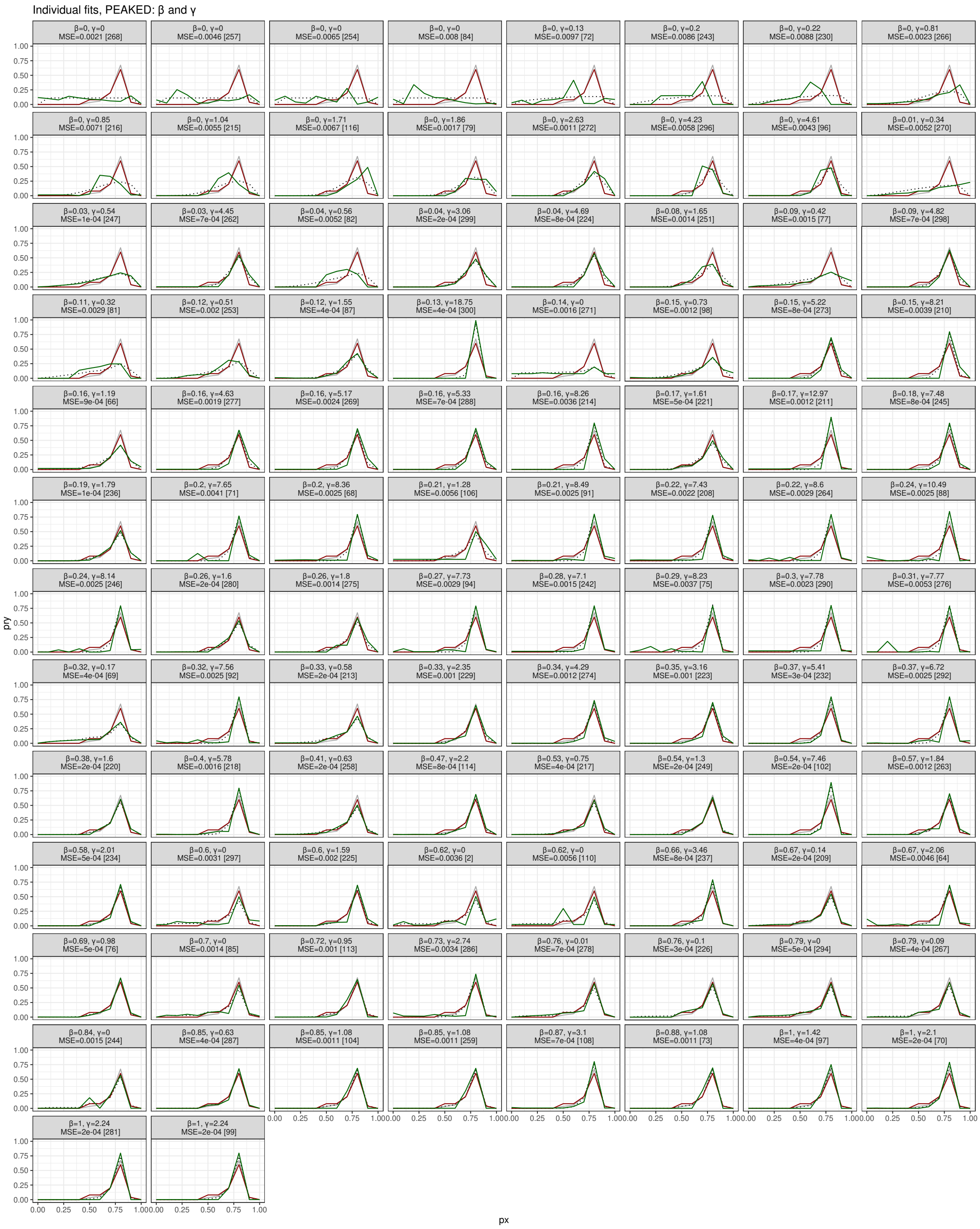
Experiment 1 BAD *participants: Each person after unlimited chips, fitting* β *and* γ



Note. Each bad individual participant after receiving unlimited chips. The red line is the reported prior, the light blue line is reported posterior, the grey line is the posterior obtained by a well-calibrated Bayesian reasoner with that prior and $\beta = \gamma = 1$, and the dotted black line is the posterior obtained by the best-fit values of β and γ for that person. The grey label for each panel contains those values as well as the mean squared error of the fit (MSE, with 0 being perfect). The number in square brackets is the participant ID.

Figure 23

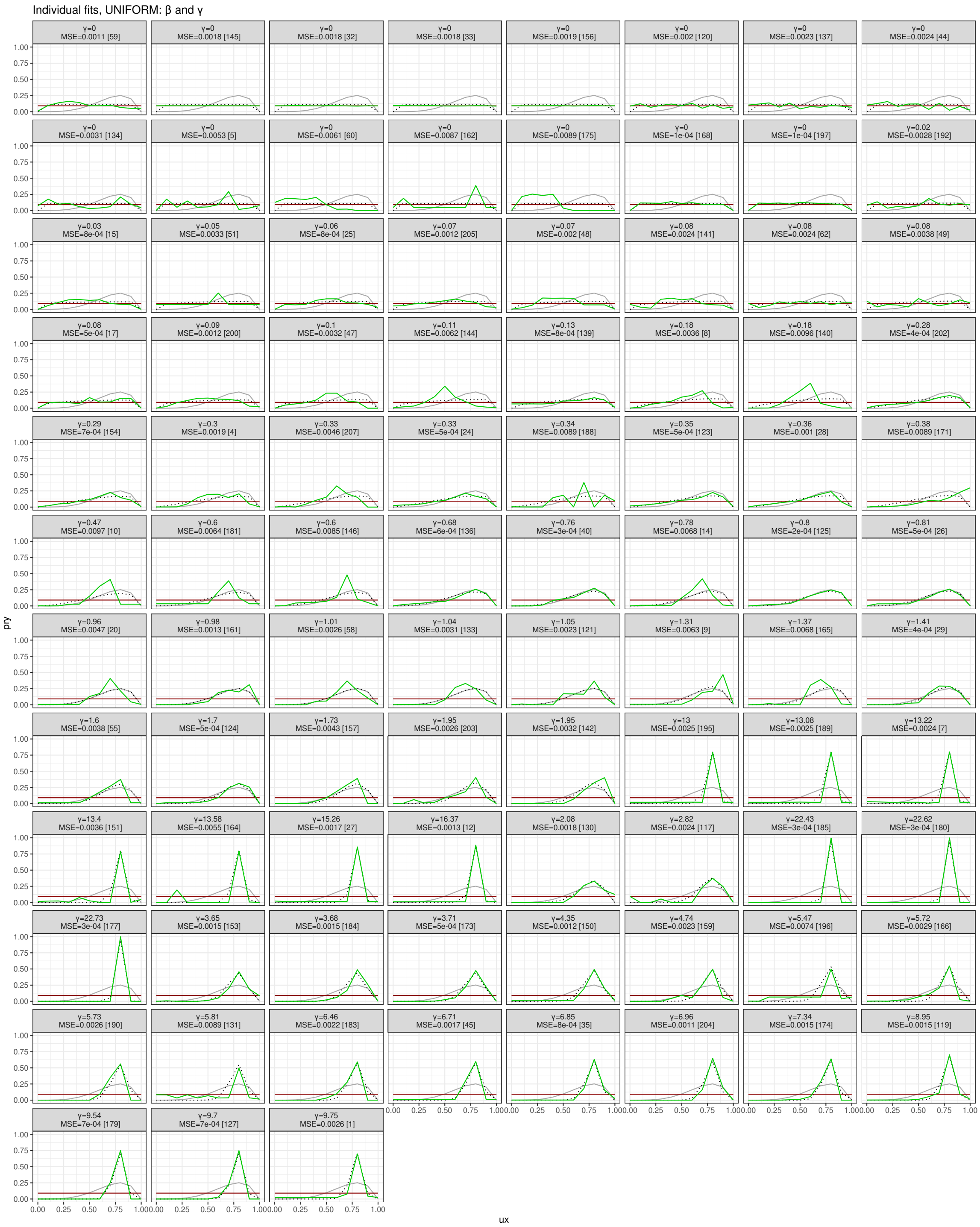
Experiment 2 GOOD *participants: Each person in the* PEAKED *condition, fitting* β *and* γ



Note. Each good individual participant in the PEAKED condition. The red line is the reported prior, the dark green line is reported posterior, the grey line is the posterior obtained by a well-calibrated Bayesian reasoner with that prior and $\beta = \gamma = 1$, and the dotted black line is the posterior obtained by the best-fit values of β and γ for that person. The grey label for each panel contains those values as well as the mean squared error of the fit (MSE, with 0 being perfect). The number in square brackets is the participant ID.

Figure 24

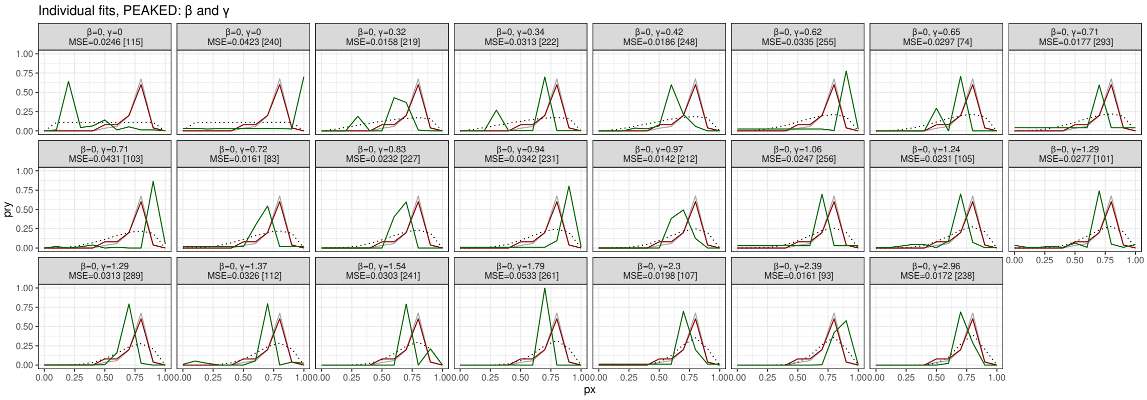
Experiment 2 GOOD *participants: Each person in the* UNIFORM *condition, fitting only* γ



Note. Each good individual participant in the UNIFORM condition. The red line is the reported prior, the light green line is reported posterior, the grey line is the posterior obtained by a well-calibrated Bayesian reasoner with that prior and $\beta = \gamma = 1$, and the dotted black line is the posterior obtained by the best-fit values of β and γ for that person. The grey label for each panel contains those values as well as the mean squared error of the fit (MSE, with 0 being perfect). The number in square brackets is the participant ID.

Figure 25

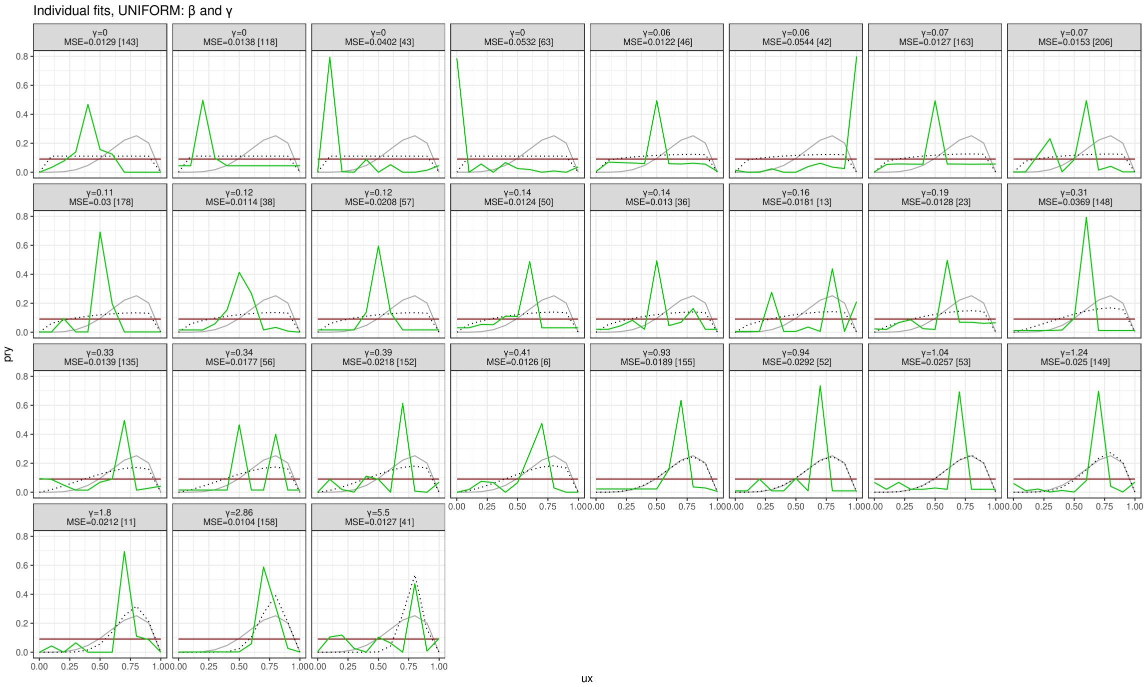
Experiment 2 BAD *participants: Each person in the* PEAKED *condition, fitting* β *and* γ



Note. Each bad individual participant in the PEAKED condition. The red line is the reported prior, the dark green line is reported posterior, the grey line is the posterior obtained by a well-calibrated Bayesian reasoner with that prior and $\beta = \gamma = 1$, and the dotted black line is the posterior obtained by the best-fit values of β and γ for that person. The grey label for each panel contains those values as well as the mean squared error of the fit (MSE, with 0 being perfect). The number in square brackets is the participant ID.

Figure 26

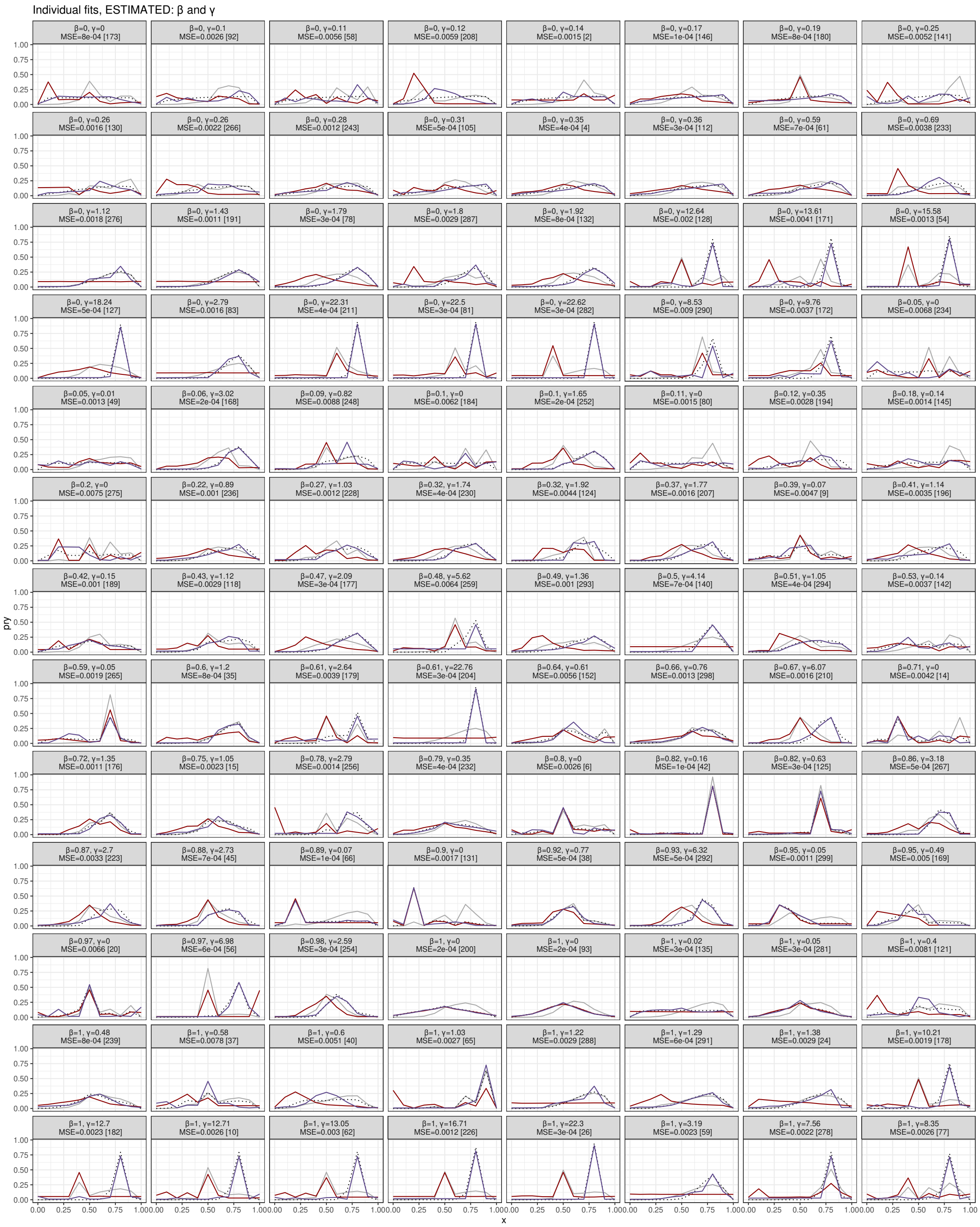
Experiment 2 BAD *participants: Each person in the* UNIFORM *condition, fitting only* γ



Note. Each bad individual participant in the UNIFORM condition. The red line is the reported prior, the light green line is reported posterior, the grey line is the posterior obtained by a well-calibrated Bayesian reasoner with that prior and $\beta = \gamma = 1$, and the dotted black line is the posterior obtained by the best-fit values of β and γ for that person. The grey label for each panel contains those values as well as the mean squared error of the fit (MSE, with 0 being perfect). The number in square brackets is the participant ID.

Figure 27

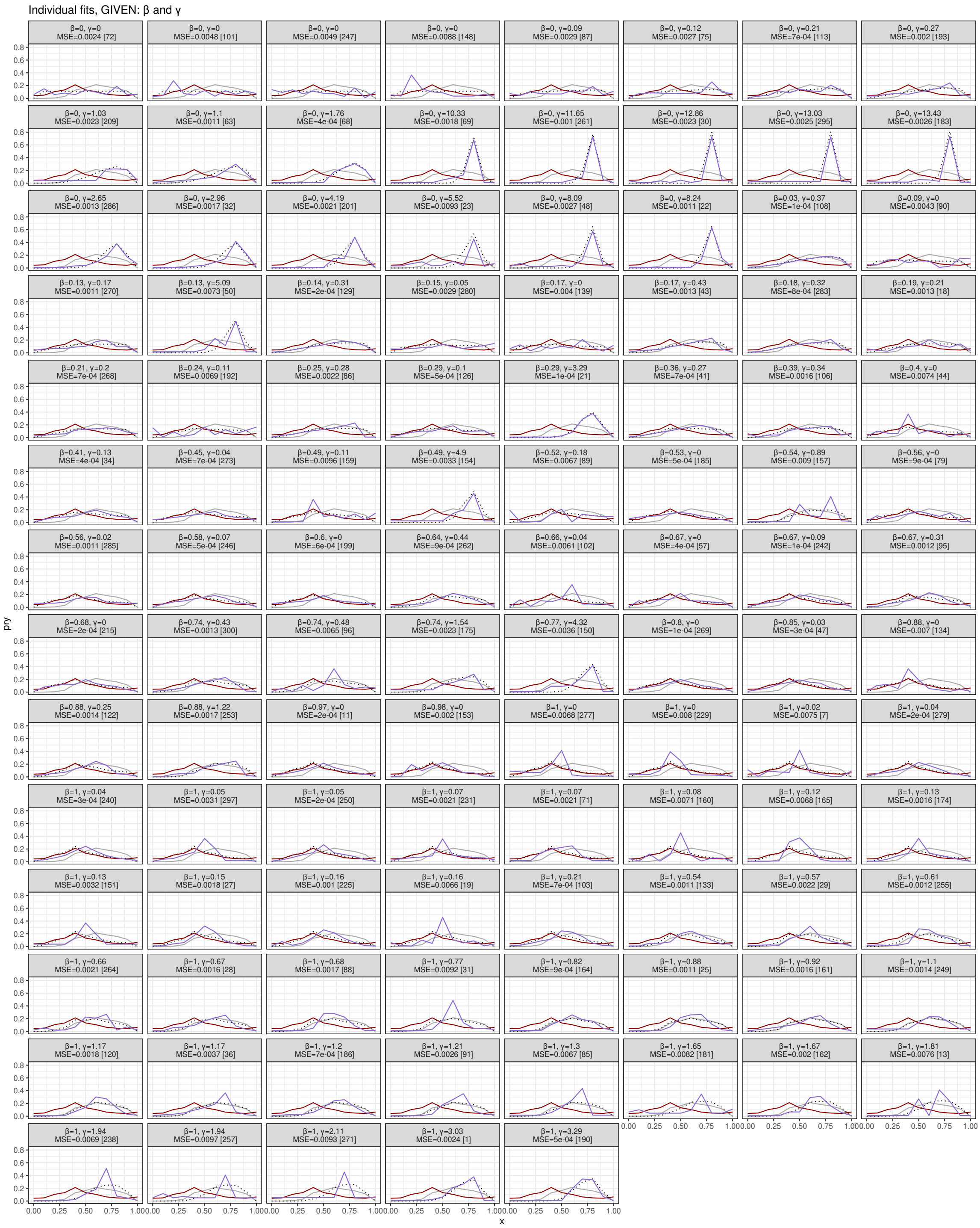
Experiment 3 GOOD *participants: Each person in the* ESTIMATED *condition, fitting* β *and* γ



Note. Each good individual participant in the ESTIMATED condition. The red line is the reported prior, the dark purple line is the reported posterior, the grey line is the posterior obtained by a well-calibrated Bayesian reasoner with that prior and $\beta = \gamma = 1$, and the dotted black line is the posterior obtained by the best-fit values of β and γ for that person. The grey label for each panel contains those values as well as the mean squared error of the fit (MSE, with 0 being perfect). The number in square brackets is the participant ID.

Figure 28

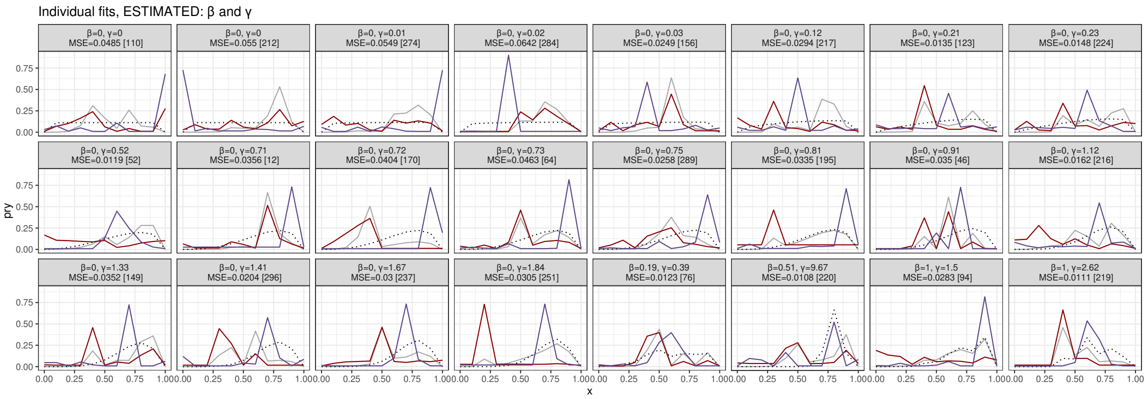
Experiment 3 GOOD *participants: Each person in the* GIVEN *condition, fitting β and γ*



Note. Each good individual participant in the GIVEN condition. The red line is the reported prior, the light purple line is the reported posterior, the grey line is the posterior obtained by a well-calibrated Bayesian reasoner with that prior and $\beta = \gamma = 1$, and the dotted black line is the posterior obtained by the best-fit values of β and γ for that person. The grey label for each panel contains those values as well as the mean squared error of the fit (MSE, with 0 being perfect). The number in square brackets is the participant ID.

Figure 29

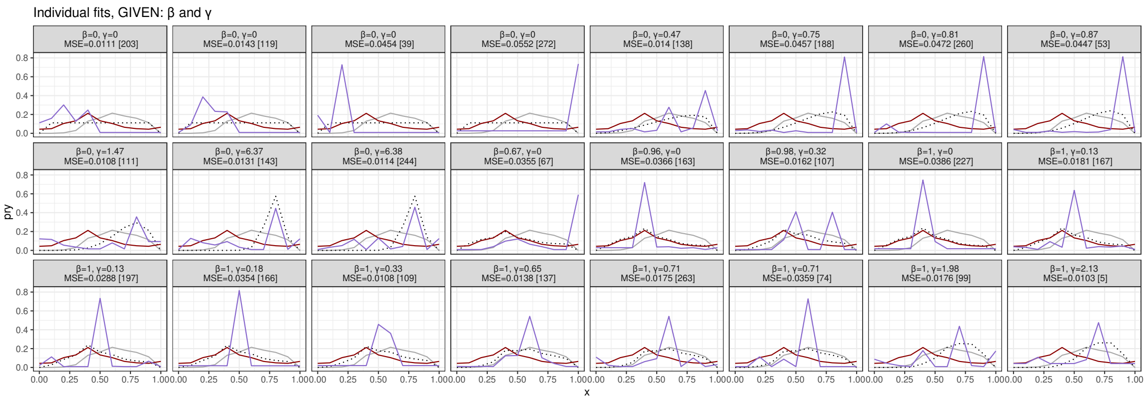
Experiment 3 BAD *participants: Each person in the* ESTIMATED *condition, fitting* β *and* γ



Note. Each bad individual participant in the ESTIMATED condition. The red line is the reported prior, the dark purple line is the reported posterior, the grey line is the posterior obtained by a well-calibrated Bayesian reasoner with that prior and $\beta = \gamma = 1$, and the dotted black line is the posterior obtained by the best-fit values of β and γ for that person. The grey label for each panel contains those values as well as the mean squared error of the fit (MSE, with 0 being perfect). The number in square brackets is the participant ID.

Figure 30

Experiment 3 BAD *participants: Each person in the* GIVEN *condition, fitting* β *and* γ



Note. Each bad individual participant in the GIVEN condition. The red line is the reported prior, the light purple line is the reported posterior, the grey line is the posterior obtained by a well-calibrated Bayesian reasoner with that prior and $\beta = \gamma = 1$, and the dotted black line is the posterior obtained by the best-fit values of β and γ for that person. The grey label for each panel contains those values as well as the mean squared error of the fit (MSE, with 0 being perfect). The number in square brackets is the participant ID.