Supplementary Material for "Heterogeneity of Rules in Bayesian Reasoning: A Toolbox Analysis"

Jan K. Woike*[1,2], Ralph Hertwig[1], and Gerd Gigerenzer[1]
[1]Max Planck Institute for Human Development, Center for Adaptive Rationality (ARC), Lentzeallee 94, 14195 Berlin, Germany
[2]University of Plymouth, School of Psychology, Portland Square, Plymouth PL4 8AA, UK

*Corresponding author, E-mail: woike@mpib-berlin.mpg.de

Abstract

Contents

List of Figures

List of Tables

The Supplementary Material is organized as follows:

1. We offer results for extended and additional simulations to strengthen and support our results in the main manuscript (section 1).
2. We present materials, questionnaires and sample details for Studies 4a, 4b, and 4c in the main manuscript (sections 2–4).
3. We offer some additional results for Studies 4a–c, including data on self-classification by participants, verbatim open-format responses, and relationships between rule use and statistical education, demographics, and other postquestionnaire variables (section 5).
4. We present additional results for Study 5, including dashboards for every single task in two categories (section 6).
5. We extend the list of studies to include task information (section 7).
6. We give details on all 106 tasks in the simulation (section 8).
7. We list all strategies found in the literature or in responses analyzed in Studies 4a–c (section 9).

1 Supplementary simulations

Here, we report on several supplementary simulations. First, we show that the fitting performance of the WA models can be achieved by a model that estimates the relative frequencies of individuals using specific rules in the toolbox. Second, we extend the simulations using the overlap criterion to include toolboxes with varying numbers of rules for both artificial and empirical datasets. Third, we present results on individual-level predictions for two datasets with a sufficient number of responses per participant.

1.1 The proportional toolbox model

As shown in the manuscript, the WA-model exhibits excellent predictive capabilities, when prediction quality is judged by the RMSE criterion. None of the single-strategy models based on psychologically realistic approaches was able to reach its level of performance, as participants differed in which of a set of strategies they used. It is nonetheless possible to target the mean of unobserved estimates: If the relative frequency of strategies in the training sample is predictive of the frequency of strategies in the test sample, then a weighted average of the individual strategies' predictions should be close to the mean of judgments, assuming that judgments not following on of the considered strategies do not systematically deviate in one direction.

This model can be easily implemented. Choosing the six discussed strategies as a basis (REP, BO, LS, FC, JO, Bayes), the weights $(w_1, \ldots, w_6)$ can be determined by calculating the observed relative frequencies in the training sample $(f_i, \ldots, f_6)$, via:

$$w_i = \frac{f_i}{\sum_{j=1}^{6} f_j}, \qquad i = 1, \ldots, 6 \tag{1}$$

The estimated mean prediction can then be calculated based on the six strategies' estimates $\hat{j}_i$:

$$\hat{\bar{j}} = \sum_{i=1}^{6} w_i \cdot \hat{j}_i \tag{2}$$

This is the prediction of the proportional toolbox model (abbreviated "Tb6" when it is built with the six discussed strategies).

1.1.1 Predictive quality of the proportional toolbox model. We added Tb6 to the prediction competition between strategies, together with one more strategy: the hit-the-middle strategy of predicting always 0.5 (called "50%"). This would be a sensible prediction if nothing is known about past judgments and the structure of the problems for which judgments have to be predicted, as it merely points at the middle of the probability scale. We again calculated the RMSE for these judgments, based on the actual simulated toolbox of rules, and compared it with the RMSE of the WA model.

Figure S1 shows the results for the added two models together with the original competitors.

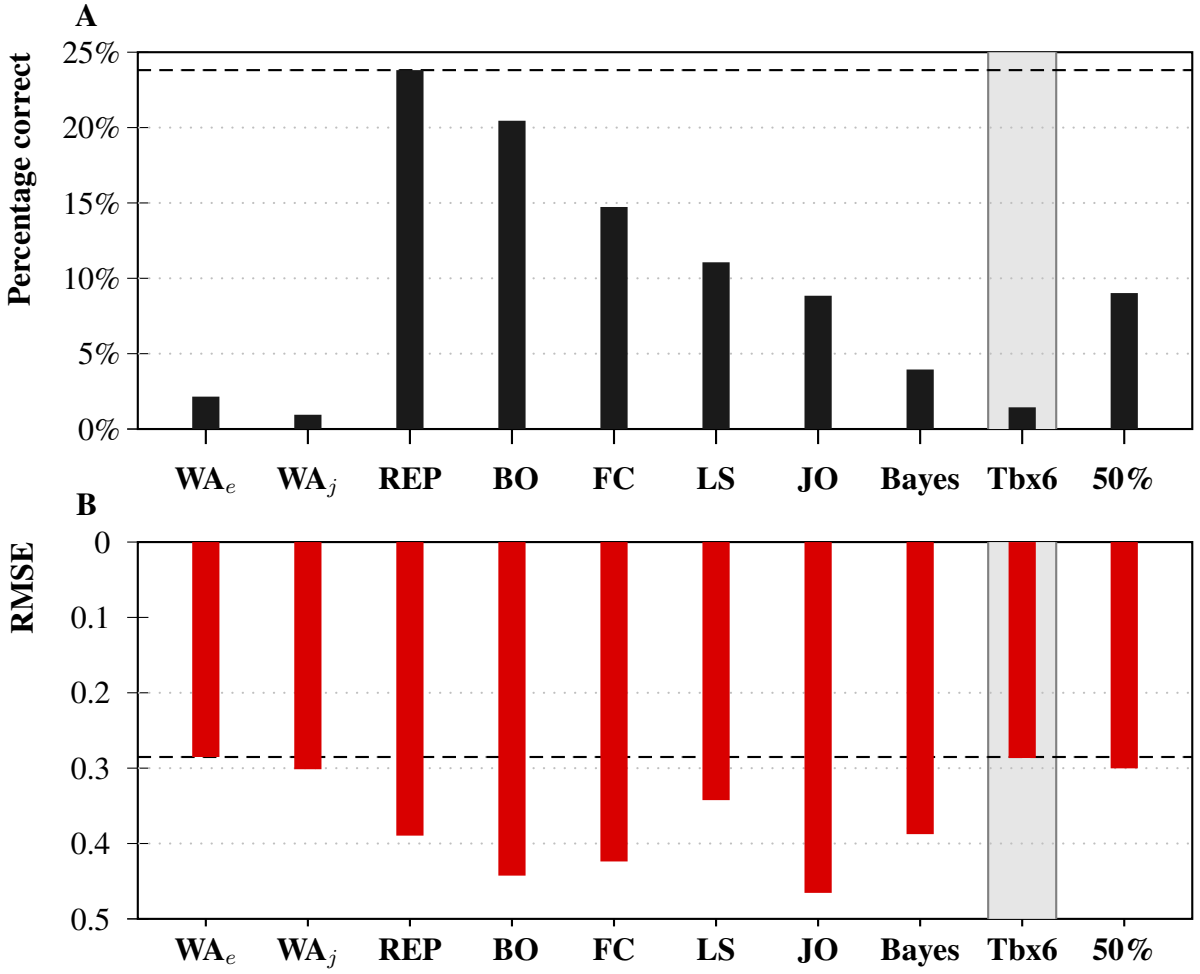

Figure S1. The Tbx6 model has an error similar to the WA$_e$ model, as measured by RMSE, for the empirical dataset of 4,188 individuals. At the same time, its percentage of correct predictions is similarly low. Shown are the average results across 10,000 trials: (A) Average percentage correct for each rule, (B) root mean squared error (RMSE) for each rule on an inverted y-axis. WA$_e$ corresponds to the empirical weighted additive model based on the data set, WA$_j$ to the weighted additive model with parameter values taken from Juslin, Nilsson, and Winman (2009). The dashed lines corresponds to the best observed performance regarding each criterion.

Although the average predicted probability judgment of the proportional toolbox model is based on the actual rules in the simulation, its RMSE is, very surprisingly, close to but still not lower than that of the 'false' WA model. This exercise shows again the enormous deceptive flexibility of this single-process model. Note though, that not a single participant is assumed to employ the proportional toolbox model, like the WA model its predictive value is solely due to its predictions falling into the middle of the judgment distribution. To bring this point home, we tested once again the"everybody estimates 50%" rule, which achieved a similarly low RMSE as the WA model.

The performance of the toolbox model was virtually indistinguishable from the WA model, both in terms of RMSE and proportion correct: the model was able to capture the mean of judgments in the test set, and like the WA model this placed the prediction apart from the actual individual judgments. The Tb6 model has two more parameters than the WA model, but the comparison via holdout set places both models on equal footing. Even more diagnostic was the performance of the parameter-free 50%-model: The model performed on a near-equal level regarding RMSE, and outperformed both the Tb6 and the WA-model regarding proportion correct: a non-negligible number of participants respond with 0.5, when facing Bayesian inference problems (the percentage seems to vary slightly between samples, but less so between problems).

1.2 Extended results for the overlap criterion

Here, we extend the simulation in the main manuscript, using the overlap criterion. We proposed in the main manuscript that the logical answer in the case of heterogeneity of processes is to abstain from making single-point predictions but predict distributions of responses. To do this, we replaced the benchmarks of RMSE and predicted-proportion by the percentage match of predicted and observed judgment distributions. To illustrate the utility of this approach, we demonstrate it here for the simulated datasets. Figure S2 shows this match both for the proportional toolbox model (based on the Five-plus model, but using only the four strategies involved in the simulation[1] and the WA model for a single split for one selected problem.

The prediction space was portioned into 101 intervals. All but two intervals covered an interval of 1% centered on percentages with natural numbers from 1% to 99%, the remaining two cover the endpoint intervals [0%; 0.5%] and [99.5%; 100%]. All model inferences that fell into the respective intervals were matched with the observed simulated inferences. The match for the toolbox model was 95.7%, relative to 0% for the WA model. Given the structure of the simulation, the match would be 100% if not for sampling error (as only half of the judgments were used to estimate the percentages). As a comparison, a model with only one generating heuristic (REP) was added to the competition: Not surprisingly, this model achieves slightly less than 25%, as it is able to capture most estimates generated by one of four strategies.

The results for this analysis involving a single Bayesian reasoning task generalize to the

[1] As the other two strategies are not involved in generating responses, the performance of the Five-plus model would be indistinguishable save for accidental overlap.

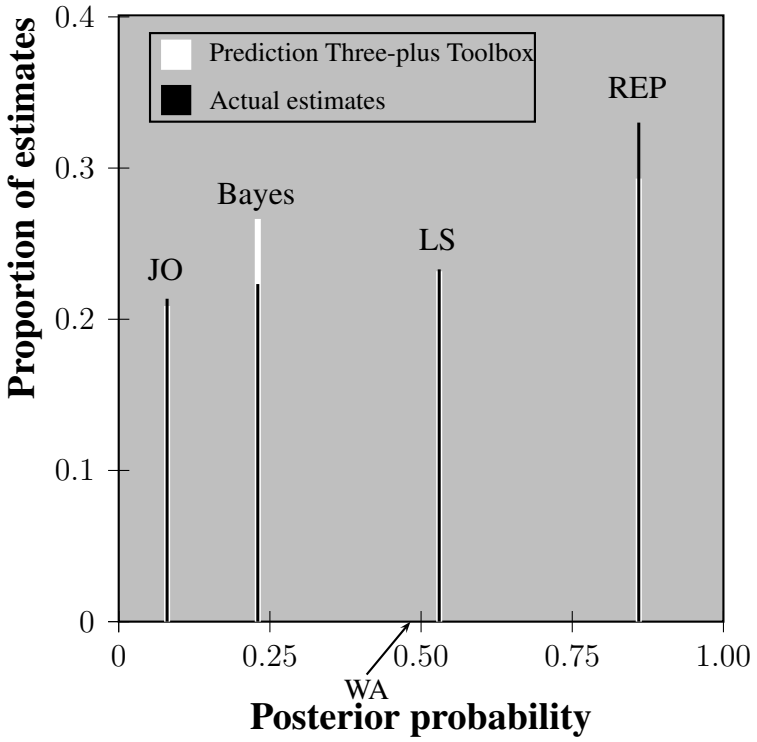

Figure S2. Predictions of the Three-Plus toolbox for a single artificial problem. The predicted proportions are shown in white bars, the empirical data is shown in black bars. The prediction of the empirical weighted additive model (WA) is also shown.

simulation involving 1,000 populations of 200 individuals responding to 40 tasks each (see Fig. S3): The toolbox distribution model clearly outperforms a single rule, and even more drastically outperforms the WA model.

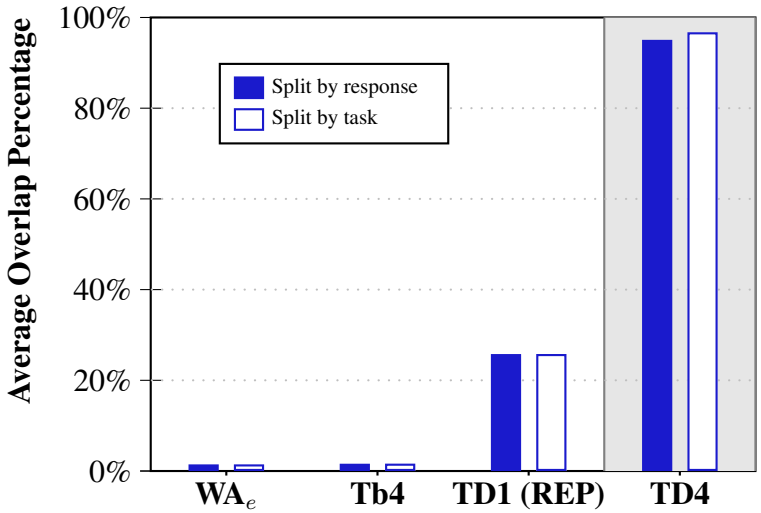

Figure S3. Average overlap percentage across 1,000 artificial populations with 100 splits for each. Splits were based on either tasks or responses. Results are shown for the weighing-and-adding model (WA), the proportional toolbox strategy (Tb4) with four rules (REP, JO, LS, Bayes), the representativeness rule (TD1(REP)), and the toolbox distribution model with the four generating non-Bayesian rules (TD4).

Figure S4 shows the same picture when the proportions and WA parameters are estimated for half of the 106 problems (or half of all responses) and predicting inferences for the other half (averaged across 10,000 different splits). Remember that we found the same qualitative results when splitting across inferences and problems. Compared to the main manuscript, we added a number of models to the competition. First, we added the proportional toolbox model to demonstrate its similarity to the WA$_e$ model. Both models are able to target mean estimates, but are of no use in making actual predictions. Second, we compared the TD6 model with more and less comprehensive models: The TD1 model only uses one heuristic (REP) and therefore performs at the same (mediocre) level as this heuristic, but is still better than the first two models. TD4 is restricted to the four heuristics used in the simulations with artificial data and achieves an intermediate performance. Adding the 50%-rule to TD6 results in the TD7 model and a slightly, but not substantially better performance.

Thus, all toolbox distribution models perform better than the simple toolbox model or the WA model with increasing performance with a higher number of rules (but only a small increase from six to seven rules). To conclude, one method to escape the problem of mistaking a highly flexible single-process theory as the appropriate theory of the process is to use a distributional performance measure. This measure compares the predicted distribution of judgments (based on the hypothesized rules) with the actual distribution rather than aggregating the distribution into one value.

One should note that the RMSE is an appropriate error measure for processes that can be characterized by a single mode with error variation. In the presence of multiple modes, it

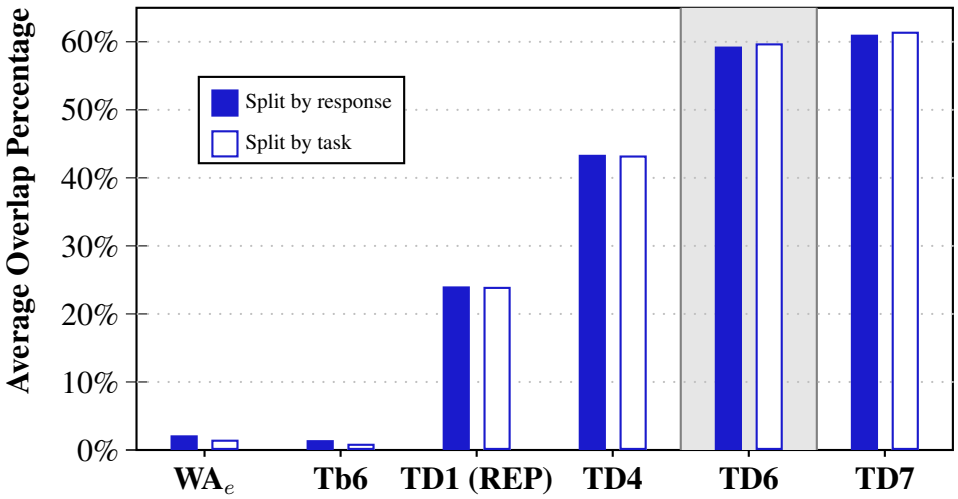

Figure S4. Average overlap percentage across 1,000 populations with 100 splits for each, based on all responses in the dataset. Splits were based on either tasks or responses. Results are shown for the weighing-and-adding model (WA), the proportional toolbox strategy (Tb6), the representativeness rule (TD1(REP)), toolbox distribution models with four (REP, JO, LS, Bayes), six (REP, BO, JO, LS, FC, Bayes), and seven (REP, BO, JO, LS, FC, Bayes, 50%) rules (TD4, TD6, TD7).

devolves into a measure of the mean of modes with a severely reduced reactivity to true errors.

1.3 Individual-level modeling

A second approach to deal with the great capacity of a flexible single-process theory is to identify the rule that an individual participant uses. Specifically, we assume that individuals use one rule in a consistent way (as shown before in Part III). We now analyze for each individual a sample of half of his or her estimates and determine the rule (out of the six rules) that is most frequently in line with those estimates. Then we use this rule to predict the individual's remaining estimates, and determine the proportions of correct predictions. Furthermore, we also analyze for each individual and using the same method, the proportion of correct predictions for the weighing-and-adding theory (with parameters estimated based on the the same half of the individual's estimates as for the tool-matching rules). This method cannot be applied across datasets and requires multiple responses per participant to make predictions testable. We tested the method with two datasets (Cohen & Staub, 2015; Juslin, Nilsson, Winman, & Lindskog, 2011) that fulfill the requirement of many responses per individual (36 and 18 responses, respectively). Figure S5A and S5C plots the average percentage correct across individuals, for each rule in the toolbox, the tool-matching model using a toolbox with the six rules, and the weighing-and-adding model, for both datasets (see Fig. S5A and S5C).

The latter model performs better than all individual rules but noticeably worse than the tool-matching model. The relatively good performance of the weighing-and-adding model is due to its ability to precisely mimic the behavior of four of the rules (see previous discussion). In fact, the weighing-and-adding model outperforms the toolbox model in terms of RMSE (see Fig. S5B and S5D). When using the parameters suggested in Juslin et al. (2009), the proportion correct drops to virtually zero and the RMSE is worse than for the matching model.

The performance of the tool-matching model is limited by two factors: (1) the model predicts well if participants follow indeed one the six strategies in the toolbox, and (2) the model predicts well if participants apply the same strategy consistently. Nonetheless, this model predicts about one in two judgments in the test set, demonstrating that many participants use simple strategies consistently.

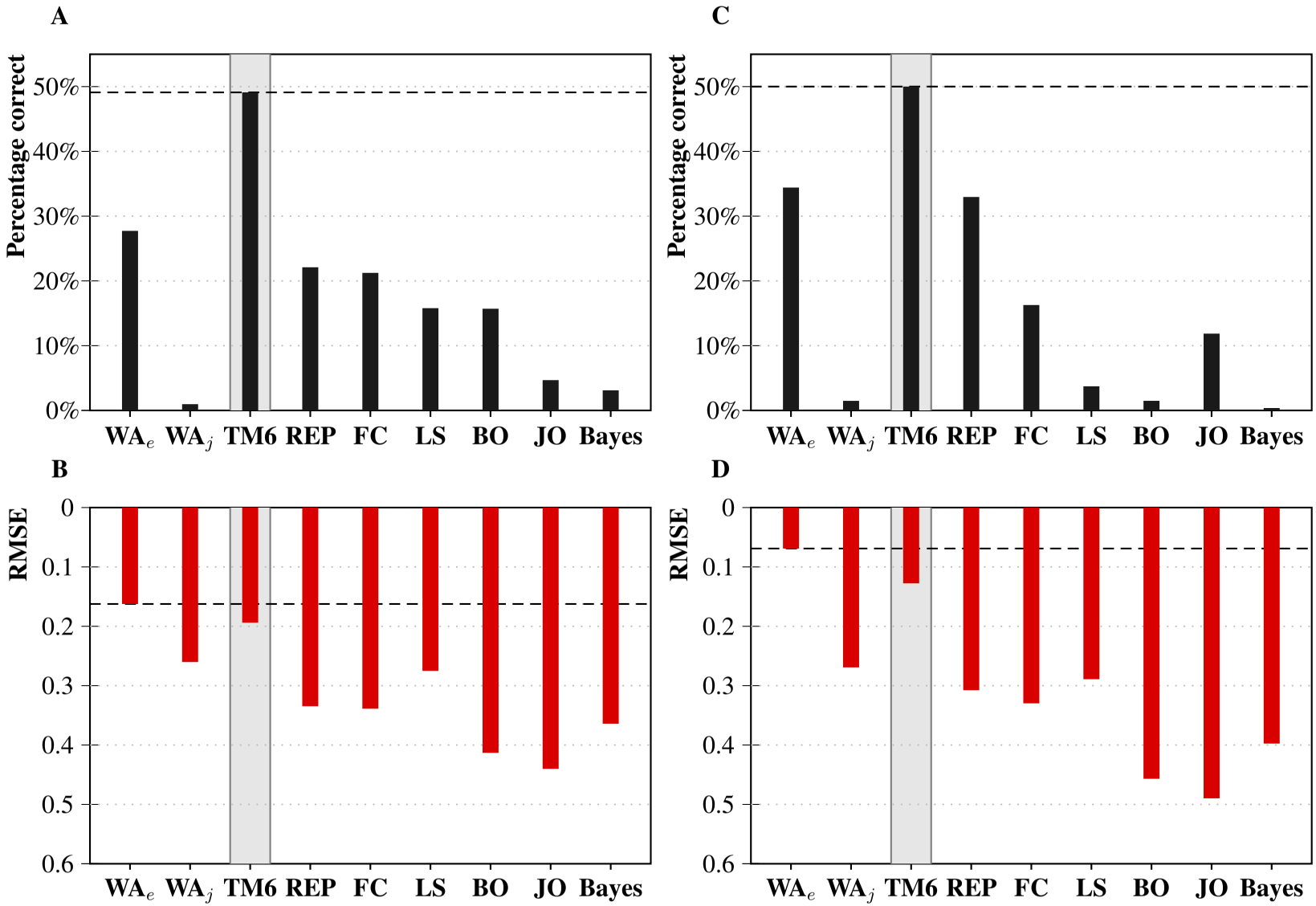

Figure S5. Results of the individual-level simulation: Results are shown for the six single rules in the toolbox on the right side, the tool-matching heuristic (TM6) in the middle, and two weighted-additive rules on the left side: WA$_e$ fits parameters to each individual, whereas WA$_j$ uses parameter values from Juslin et al. (2009). (A) Percentage correct for Cohen and Staub (2015, 36 tasks, N=95); (B) Root mean squared error (RMSE) for Cohen and Staub (2015) on an inverted y-axis; (C) Percentage correct for Juslin et al. (2011, 18 tasks, N=15); (D) RMSE for Juslin et al. (2011).

1.4 Variation of Study2 with a different set of heuristics

In this variation of Study 2, we replaced the base-rate only rule by the representativeness rule (see Table S1). These are the exact three rules considered and studied by Juslin et al. (2009, p. 869).

Table S1
Definition of Bayes's rule and three non-Bayesian rules used in the simulation, and the number of cues used by each rule. b = base rate or prior probability; h = hit rate; f = false alarm rate.

	Rule	Formula	Cues used
1)	Bayes's rule	$P(H\|D) = \frac{b \cdot h}{b \cdot h + (1-b) \cdot f}$	3
2)	Joint occurrence	$P(H\&D) = b \cdot h$	2
3)	Likelihood subtraction	$P(D\|H) - P(D\|\neg H) = h - f$	2
4)	Representativeness	$P(D\|H) = h$	1

With this set of rules, we created another 1,000 populations of 200 simulated respondents each; and again each respondent received 40 tasks with the same constraints ($h > 0.5$ and $f < 0.5$) and an equal probability for each respondent to use one of the four rules across all tasks.

Again, we demonstrate how well the four rules can be used to predict responses compared to the weighing-and-adding (WA) model using one simulated task as an example. We fitted the WA model to a full set of 40 tasks (based on half of all responses), and calculated the RMSE for this model and each of the rules actually assigned to respondents. Figure S6 summarizes this demonstration in parallel to Figure 2 in the main manuscript. Again, the WA model prediction is closest to the average response and therefore shows a smaller RMSE than each of the generating rules.

We then proceeded to the full analysis of all 1,000 populations with 100 repetitions per population, including the conservatism model (see Figure S7). Similar to the results in the main manuscript, the WA model performs best in predicting responses without being used by any respondent in the virtual populations. At the same time, the CON model performs worse than before. This can be explained by the difference in flexibility between the two parameterized models: the CN model is constrained to predict a point between the responses of Bayes' rule and the base rate. Replacing the base-rate only strategy by the representativeness rule, takes one of these flanking rules out of the simulation.

This difference is exacerbated when modelling on the individual level. Figure S8 shows the results of another simulation, mirroring the individual-level simulation in the main manuscript. We replicate the finding that the WA rule is the best model for out-of-sample predictions for both criteria. It has the smallest RMSE and the highest percentage of correct predictions. But while the WA model is able to mirror the representativeness rule ($w_b = 0$, $w_h = 1$, $w_f = -1$, $\alpha = 0$, $\epsilon = 0$), the same is not possible for the CON model. Its ability to mirror Bayes' rule ($w = 0$, $\epsilon = 0$) allows it to match the performance of the generating rules.

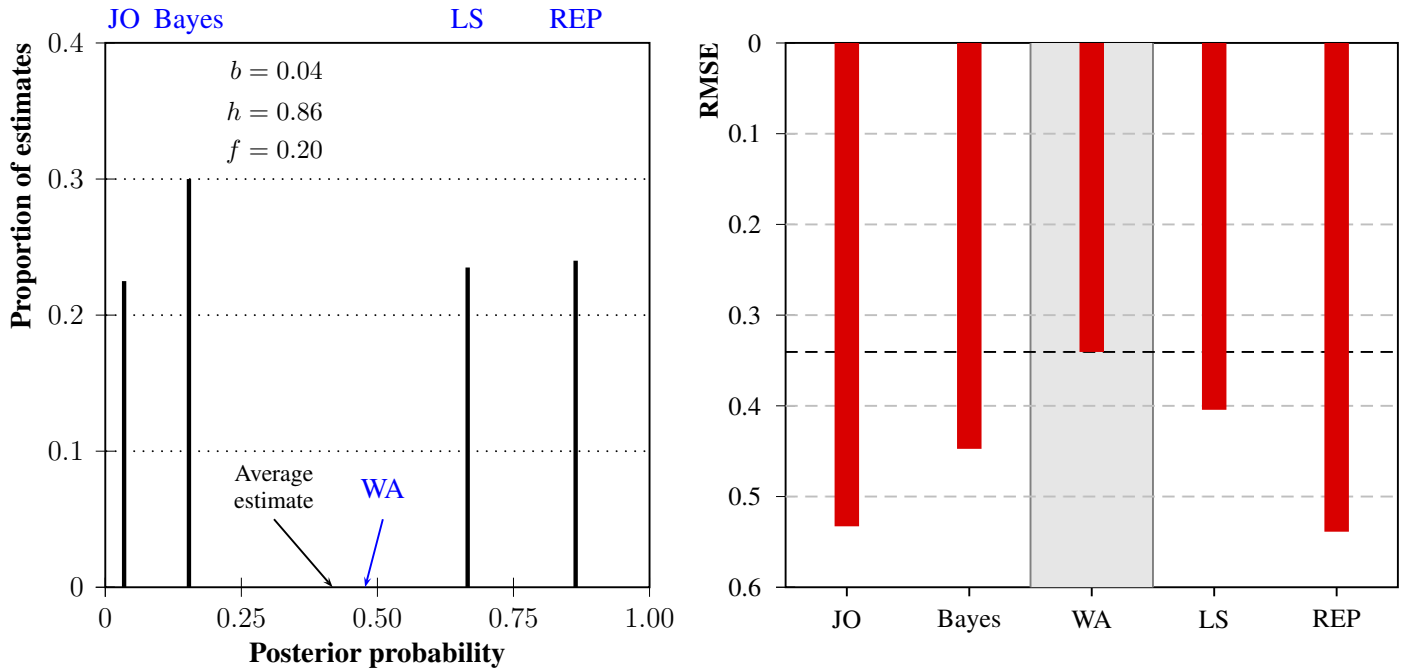

Figure S6. The weighing-and-adding model has the lowest root mean squared error (RMSE) in predicting the inferences of simulated individuals, none of whom used weighing-and-adding. Left panel: The posterior probabilities provided by 200 simulated respondents in one of the 40 generated Bayesian tasks. The proportion of estimates for each of the four generating rules is shown, as is their average. The parameters of the weighing-and-adding model (WA) were estimated based on half of the dataset; its predictions in the other half of the data set are plotted. Right panel: RMSE for the five strategies (four rules plus WA) for the same task. LS = likelihood subtraction; Bayes = Bayes's rule; REP = representativeness; JO = joint occurrence.

Nonetheless, the CON model outperforms the four strategies actually used by respondents in terms of the RMSE.

2 Details for Study 4a

The estimates for Study 4a will be found on the Harvard Dataverse (https://doi.org/10.7910/DVN/FYMODJ).

2.1 Sample

A total of 548 participants passed the initial attention checks and started Study 4a. Three of them gave incomplete responses and were not included in the analysis. Regarding the 545 participants who gave all estimates, 250 were male (45.9%), 295 female (54.1%). The age range was between 18 and 77 years, with a mean of 36.1 years ($SD = 12.72$).

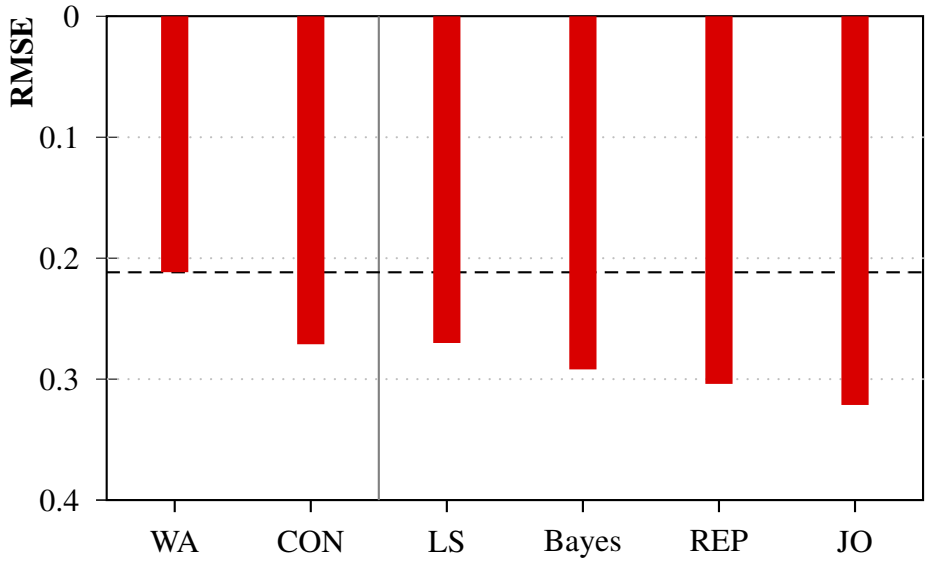

Figure S7. The out-of-sample predictive power of the weighing-and-adding (WA) model that is not employed by any of the simulated respondents. Shown is the average RMSE for each rule across 1,000 populations of 200 simulated respondents for all 40 tasks. Parameters for the WA model were estimated based on half of the responses in each population. LS =likelihood subtraction; Bayes = Bayes's rule; REP = representativeness; JO = joint occurrence.

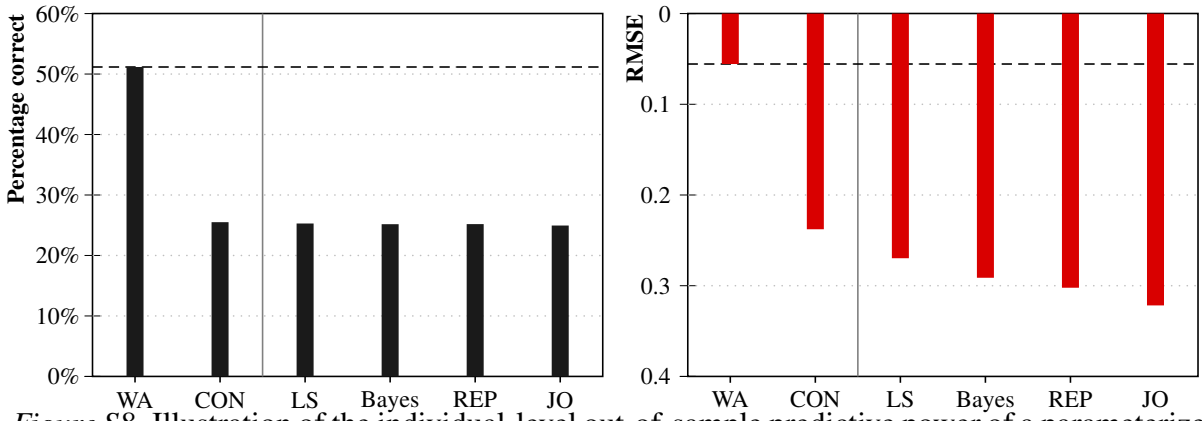

Figure S8. Illustration of the individual-level out-of-sample predictive power of a parameterized WA model. Results for the five rules across 1,000 populations of 200 simulated respondents with 40 tasks; individual parameters for the WA model and CON model were estimated based on half of the responses for each individual. The left panel shows the average proportion of correct predictions, the right panel the RMSE for each of the rules. The dashed lines mark the best observed performance for each criterion (WA in both cases).

Table S2
Tasks in Study 4a

Task	b	h	f
1	0.50	0.80	0.40
2	0.10	0.95	0.50
3	0.20	0.30	0.10
4	0.60	0.50	0.25

2.1.1 MTurk Specification.

- Time estimate: 10–20 minutes
- Fixed payment: $1.25
- Variable additional payment: up to $1.00 (on average about $0.50)
- Filters: HIT approval rating $\geq 95\%$, Number of approved HITs >50, Location: US
- Attention checks: Participants had to pass at least one of two consecutive attention checks at the beginning of the study

2.1.2 MTurk Context. For Study 4a and 4b, participants completed three study parts before the Bayes section of the survey. All of these tasks were economic games, in the first and third block participants played one of two possible games. The games in the first block consisted of task involving money burning (Zizzo, 2003) or a promising task in which participants were offered sums of money for a promise to pay them back at the end of the study (Woike & Kanngiesser, 2019). The second part consisted of an allocation task testing principles of distributive justice, similar to Engelmann and Strobel (2004). The third task involved a variant of Newcombe's problem (Nozick, 1969) or a variant of the ultimatum game (Güth, Schmidt, & Sutter, 2007).

2.2 Method

Each participant made four estimates, one for each scenario and (simultaneously) one for each set of probabilities shown in Table S2. Probabilities were randomly assigned to tasks for each participant, with pairings cross-balanced. The scenarios are described in the next section, with placeholders [BR] for base rate b, [HR] for hit rate h and [FA] for false alarm rate f. Scenarios were chosen to vary from social and personal settings to an abstract numeric task.

2.3 Scenarios

2.3.1 Instructions. In the following, you will be faced with four scenarios involving probabilities. Those questions have correct answers and your responses will be compared to the correct answers to determine your payment for this part.

[*New Survey Page*]

2.3.2 Scenario 1: Game playing. *(See Figure S9 for a screenshot of this question.)*

You are playing a game of poker against an opponent who you have studied well over several years. You have noticed that he sometimes blinks rapidly when he decides to bet against you.

Overall you noticed that he had a probability of [BR] to win against you when he bets against you.

You also noticed that in those games he bet against you and won, he blinked with a probability of [HR].

You also noticed that in those games when he bet against you and lost he blinked with a probability of [FA].

In this game you see that your opponent blinked when he bet against you.

What is the probability that he will win?

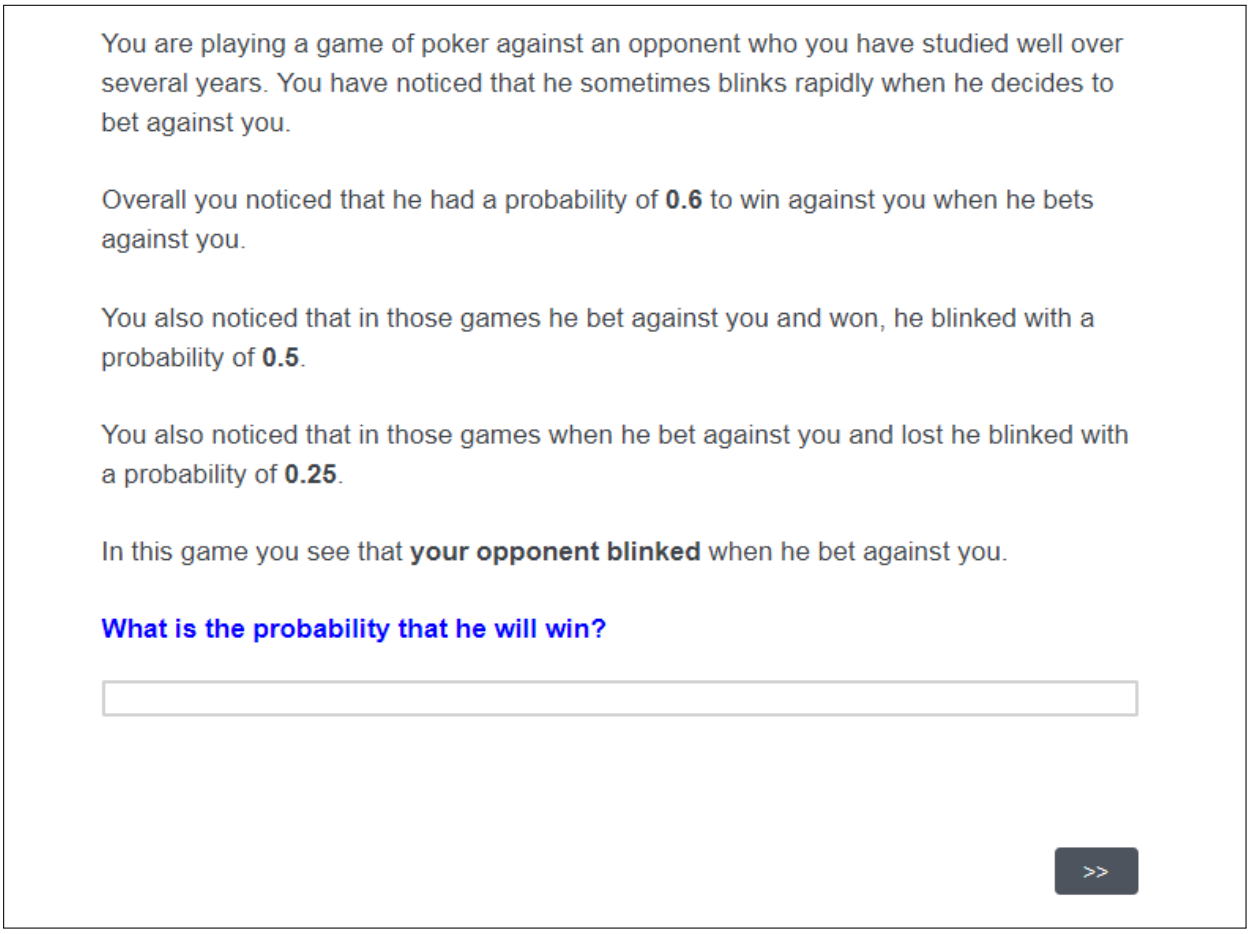

Figure S9. Inference in Study 4a

2.3.3 Scenario 2: Social prediction. You are having dinner with a friend, who you have known for many years.

Over the many times you went out to have dinner, your friend enjoyed the meal with a probability of [BR].

You noticed that in those cases in which he enjoyed the meal, he ate very fast with a probability of [HR].

You also noticed that in those cases in which he did not enjoy the meal, your friend ate very fast with a probability of [FA].

This time you observe that your friend eats very fast.

What is the probability that your friend enjoys the meal?

2.3.4 Scenario 3: Personalized car. You are driving a car that you know very well.

You know that you will encounter a technical problem on a drive with a probability of [BR].

You noticed that in those cases in which your encountered a technical problem, you observed a warning signal when starting the engine with a probability of [HR].

You also noticed that in those cases in which you did not encounter any technical problem, you observed a warning signal when starting the engine with a probability of [FA].

This time you observe a warning signal when starting the engine.

What is the probability that you will encounter a technical problem?

2.3.5 Scenario 4: Abstract urn. You are drawing objects from an urn.

You know that you will draw a blue object with a probability of [BR].

You noticed that in those cases in which you drew a blue object, the object was round with a probability of [HR].

You also noticed that in those cases in which you did not draw a blue object, the object was round with a probability of [FA].

This time you know that the object you drew is round.

What is the probability that the object you drew is blue?

2.4 Postquestionnaire

In a postquestionnaire spanning several pages, participants were asked about the strategy they used for making judgments in the preceding tasks.

2.4.1 Comparison with others. First, participants were asked to judge their own performance in comparison to other participants, by sorting themselves subjectively into a quartile based on the number of correct responses they had given. See Figure S10 for a screenshot of this question.

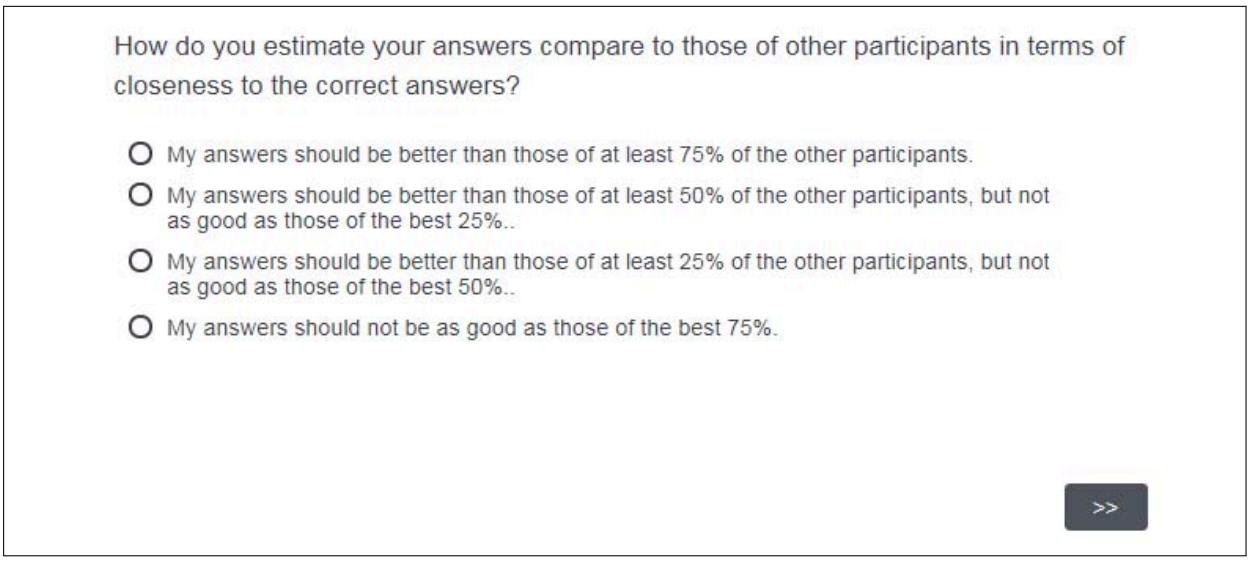

Figure S10. Postquestionnaire: Comparison with others

2.4.2 Cue use. Participants were given a general scheme of the four previous tasks, defining the terms base rate, hit rate, and false alarm rate. They were then asked, which of the three pieces of information they utilized when making judgments. See Figure S11 for a screenshot of this question.

The previous tasks all had a similar structure.

You were informed about how likely a specific attribute is in general (*for example the probability that any member of a population has a disease*). This is also called the Base Rate (**BR**).

You were given diagnostic information about a case (for example the result of a diagnostic test procedure).

a) You were told how likely a positive result is if the attribute is present (*the test has a positive result for a person with a disease*). This is also called the Hit Rate (**HR**).

b) You were told how likely a positive result is if the attribute is **not** present (*the test has a positive result for a healthy person*). This is also called the False Alarm Rate (**FA**).

You are told that a test result is positive for a given case and you were asked to calculate the probability that this case has the attribute (*for example: what is the probability that a person with a positive test result has the disease?*).

Which of these three values did you use to come up with a answer for the previous scenarios (*check all that apply, 0-3 checks*)?

- [ ] the Base Rate (**BR**)
- [x] the Hit Rate (**HR**)
- [ ] the False Alarm Rate (**FA**)

>>

Figure S11. Postquestionnaire: Cue use

2.4.3 Strategy use. On the following page, participants again saw the general scheme with definitions, and were then asked to identify their chosen strategy among a set of options (including the option to enter non-listed strategies). To allow for participants' switching strategies, more than one strategy could be chosen from the list. We wanted to exclude the possibility that participants could identify the normative theory by mere complexity, thus we included distractors in the strategy list. The list covered the following strategies:

- Unsystematic
- BO: base rate only, b
- REP: representativeness, h
- FO: false alarm rate only[2], f
- LS: likelihood subtraction, $h - f$
- JO: joint occurrence, $b \cdot h$
- WA: weighting of h, b, f
- Bayes: Bayes' theorem, $\frac{b \cdot h}{b \cdot h + (1-b) \cdot f}$
- FC: false alarm complement, $1 - f$
- Adding the positive, subtracting the negative: $b + h - f$
- Distractor: $\frac{b \cdot (b+h)}{b \cdot h - f \cdot (1-b)}$
- Individual systematic approach (open-format answer)
- Individual unsystematic approach (open-format answer)

See Figure S12 for a screenshot of this question.

[2]This implies a*positive* correlation between false alarm rate and posterior.

The previous tasks all had a similar structure.

You were informed about how likely a specific attribute is in general (*for example the probability that any member of a population has a disease*). This is also called the Base Rate (**BR**).

You were given diagnostic information about a case (for example the result of a diagnostic test procedure).

a) You were told how likely a positive result is if the attribute is present (*the test has a positive result for a person with a disease*). This is also called the Hit Rate (**HR**).

b) You were told how likely a positive result is if the attribute is **not** present (*the test has a positive result for a healthy person*). This is also called the False Alarm Rate (**FA**).

You are told that a test result is positive for a given case and you were asked to calculate the probability that this case has the attribute (*for example: what is the probability that a person with a positive test result has the disease?*).

What did you do to come up with an answer for the previous scenarios? (*check all that apply, at least one option*)

- ☐ I answered with the first number that came to my mind, there was no systematic principle I used.
- ☐ I answered by choosing a number close to the Base Rate (**BR**).
- ☐ I answered by choosing a number close the Hit Rate (**HR**).
- ☐ I answered by choosing a number close to the False Alarm Rate (**FA**).
- ☐ I calculated: **HR** - **FA** or a number close to it.
- ☐ I calculated: **BR** x **HR** or a number close to it.
- ☐ I multiplied **BR**, **HR**, and **FA** with specific weights and added the results. *(explain if you want)*
- ☐ I calculated: (**BR** x **HR**) / (**BR** x **HR** + (1- **BR**) x **FA**) or a number close to it.
- ☐ I calculated: 1-**FA** or a number close to it.
- ☐ I calculated: **BR** + **HR** - **FA** or a number close to it.
- ☐ I calculated: (**BR** x (**BR** + **HR**))/ (**BR** x **HR**- **FA** x (1-**BR**)) or a number close to it.
- ☐ I used a different systematic approach (*please explain*)
- ☐ I used a different approach but it was not systematic (*please explain*)

Figure S12. Postquestionnaire: Strategy use

2.4.4 Statistical self-assessment. Finally, participants were asked about their previous experience with the statistical concepts underlying the normative response. One question addressed the theorem itself, a second question asked for previous classes in statistics or probability theory and a third asked for the attitude towards tasks involving calculations. See Figure S13 for a screenshot of this question.

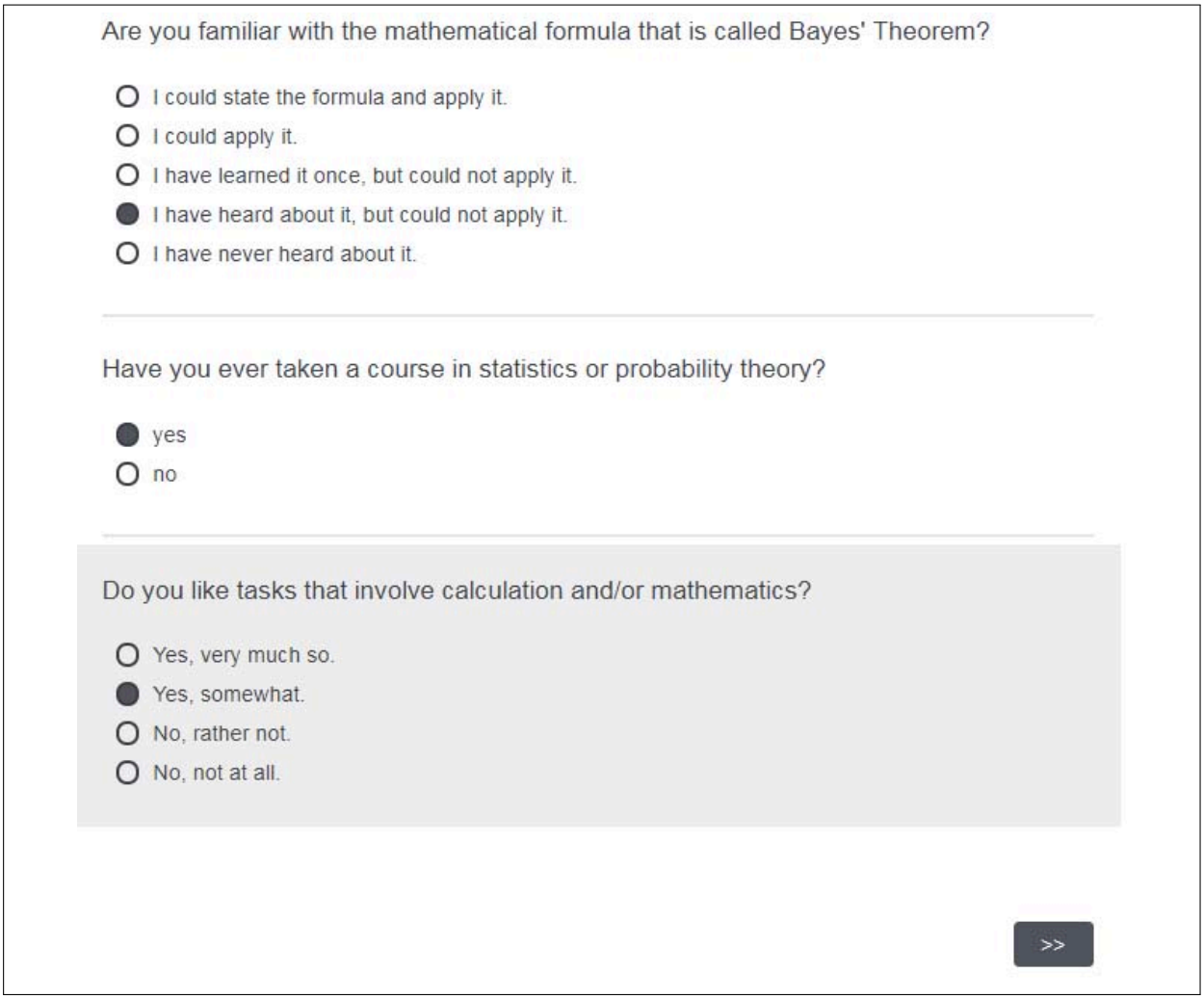

Figure S13. Postquestionnaire: Statistical self-assessment

2.4.5 Open-format answers. Participants were given a comment field to enter any comment or idea they were not able to express before. These comment fields, together with open-format answers describing strategies, were used as the basis for identifying strategies for the fitting competition.

3 Details for Study 4b

The estimates for Study 4b will be found on the Harvard Dataverse (https://doi.org/10.7910/DVN/FYMODJ).

3.1 Sample

The data for Study 4b was collected in parallel with Study 4a. Participants were randomly assigned to Study 4a or 4b with an intended proportion of 5:3. A total of 328 participants passed the initial attention checks and started Study 4b, ten participants gave incomplete responses, so that 318 participants formed the final sample size. Regarding the 318 participants who gave all estimates, 141 were male (44.3%), 177 female (55.7%). The age range was between 18 and 79 years (one invalid age entry), with a mean of 36.6 years ($SD = 12.48$).

3.1.1 MTurk Specification.

- Time estimate: 10–20 minutes
- Fixed payment: $1.25
- Variable additional payment: up to $1.00 (on average about $0.50)
- Filters: HIT approval rating $\geq 95\%$, Number of approved HITs >50, Location: US
- Attention checks: Participants had to pass at least one of two consecutive attention checks at the beginning of the study

3.1.2 MTurk Context.

See the description for Study 4a.

3.2 Method

Participants in Study 4b faced two problems each. Again, probabilities were separated from scenarios. Each participants saw one problem with the probabilities for Task 1a or Task 1b, and another problem with the probabilities for Task 2a or 2b in Table S3. The assignment of the randomly chosen probabilities to scenarios was again random.

Participants responded to one scenario in a health setting and one scenario in a factory production setting. Each scenario was chosen out of four possible variants, varying the two factors of mild vs. severe consequences, and main focus on individual vs. numbers.

These factors correspond to two of the three qualitative factors analyzed by Hafenbrädl and Hoffrage (2015) in an analysis of 19 tasks (i.e., stakes and focus). Note that the dataset entered into the analyses in the main manuscript clustered responses based on the probabilities alone (ignoring differences in scenarios).

Table S3
Tasks in Study 4b

Task	b	h	f
1a	0.20	0.90	0.05
1b	0.20	0.90	0.25
2a	0.01	0.99	0.20
2b	0.001	0.999	0.20

3.3 Scenarios

Scenarios 1.1 to 1.4 form the first block with scenarios situated in the health domain. Scenarios 2.1 to 2.4 form the second block with scenarios situated in a factory setting.

3.3.1 General Introduction. In the following, you will be faced with two scenarios involving probabilities. Those questions have correct answers and your responses will be compared to the correct answers to determine your payment.

3.3.2 Scenario 1.1: Genetic condition, mild consequences, individualized. *(See Figure S14 for a screenshot of this question.)*

Imagine that there is a rare genetic condition in the population. A person with this condition will go through life without symptoms until they reach an age of about 40 years.

At this stage, the condition will cause affected persons to suffer from periods of sneezing and mild coughs for the rest of their lives. Otherwise the condition is benign and there is no change in life expectancy between persons with or without this condition.

The probability to have this genetic condition in the population is [BR].

A test procedure has been researched to determine whether a person has or does not have the condition. If a person has the condition, the test will be positive with a probability of [HR].

If a person does not have the condition, the test is still positive with a probability of [FA].

Now imagine that a dear friend of yours underwent the test procedure. The result of the test came back positive. Your friend asks you how this result can be interpreted. Your friend asks you how this result can be interpreted and would like to understand what to expect.

What is the probability that your friend has the genetic condition?

3.3.3 Scenario 1.2: Genetic condition, mild consequences, statistical. Imagine that there is a rare genetic condition in the population. A person with this condition will go through life without symptoms until they reach an age of about 40 years.

At this stage, the condition will cause affected persons to suffer from periods of sneezing and mild coughs for the rest of their lives. Otherwise the condition is benign and there is no change in life expectancy between persons with or without this condition.

The probability to have this genetic condition in the population is [BR].

A test procedure has been researched to determine whether a person has or does not have the condition. If a person has the condition, the test will be positive with a probability of [HR].

If a person does not have the condition, the test is still positive with a probability of [FA].

Imagine that there is a rare genetic condition in the population. A person with this condition will go through life without symptoms until they reach an age of about 40 years.

At this stage, the condition will cause affected persons to suffer from periods of sneezing and mild coughs for the rest of their lives. Otherwise the condition is benign and there is no change in life expectancy between persons with or without this condition.

The probability to have this genetic condition in the population is **0.2**.

A test procedure has been researched to determine whether a person has or does not have the condition. If a person has the condition, the test will be positive with a probability of **0.9**.

If a person does not have the condition, the test is still positive with a probability of **0.05**.

Now imagine that a dear friend of yours underwent the test procedure. The result of the test came back positive. Your friend asks you how this result can be interpreted. Your friend asks you how this result can be interpreted and would like to understand what to expect.

What is the probability that your friend has the genetic condition?

>>

Figure S14. Inference in Study 4b

Now imagine that a hospital tries to establish procedures for dealing with the group of patients that test positive on this test. The hospital would like to enable doctors to help patients to interpret test results and to understand what to expect in the future. You are consulted in this process and asked a question:

What is the probability that a member of the group that tests positive has the genetic condition?

3.3.4 Scenario 1.3: Genetic condition, severe consequences, statistical. Imagine that there is a rare genetic condition in the population. A person with this condition will go through life without symptoms until they reach an age of about 40 years.

At this stage, the condition will cause affected persons to suffer from periods of severe fever and organic malfunctions for the rest of their lives. The life expectancy of affected persons is reduced by 20 years compared to persons without this condition.

The probability to have this genetic condition in the population is [BR].

A test procedure has been researched to determine whether a person has or does not have the condition. If a person has the condition, the test will be positive with a probability of [HR].

If a person does not have the condition, the test is still positive with a probability of [FA].

Now imagine that a hospital tries to establish procedures for dealing with the group of patients that test positive on this test. The hospital would like to enable doctors to help patients to interpret test results and to understand what to expect in the future. You are consulted in this process and asked a question:

What is the probability that a member of the group that tests positive has the genetic condition?

3.3.5 Scenario 1.4: Genetic condition, severe consequences, individualized. Imagine that there is a rare genetic condition in the population. A person with this condition will go through life without symptoms until they reach an age of about 40 years.

At this stage, the condition will cause affected persons to suffer from periods of severe fever and organic malfunctions for the rest of their lives. The life expectancy of affected persons is reduced by 20 years compared to persons without this condition.

The probability to have this genetic condition in the population is [BR].

A test procedure has been researched to determine whether a person has or does not have the condition. If a person has the condition, the test will be positive with a probability of [HR].

If a person does not have the condition, the test is still positive with a probability of [FA].

Now imagine that a dear friend of yours underwent the test procedure. The result of the test came back positive. Your friend asks you how this result can be interpreted and would like to understand what to expect.

What is the probability that your friend has the genetic condition?

3.3.6 Scenario 2.1: Factory production, mild consequences, individualized. Imagine that there is a factory producing cars that is concerned about a specific fault in the production process. A car with this fault will show no problems until it reaches a mileage of about 3000 miles.

At this point, the production fault will cause affected cars to undergo a change of coloring in small areas (of about one square inch) on the backside of the car. This color change is only cosmetic in nature does not affect the safety properties of the car.

The probability to have this production fault in any car coming out of the production line is [BR].

A quality control routine has been implemented to check whether a has the production fault or not. If a car has the fault, the routine will flag this car with a probability of [HR].

If a car does not have the condition, the routine will still flag the car with a probability of [FA].

Now imagine that you have been offered a car from this factory at a discount price. You obtain the information that the car was flagged by the quality control routine. You wonder what to expect about this car.

What is the probability that this car has the production fault?

3.3.7 Scenario 2.2: Factory, mild consequences, statistical. Imagine that there is a factory producing cars that is concerned about a specific fault in the production process. A car with this fault will show no problems until it reaches a mileage of about 3000 miles.

At this point, the production fault will cause affected cars to undergo a change of coloring in small areas (of about one square inch) on the backside of the car. This color change is only cosmetic in nature does not affect the safety properties of the car.

The probability to have this production fault in any car coming out of the production line is [BR].

A quality control routine has been implemented to check whether a has the production fault or not. If a car has the fault, the routine will flag this car with a probability of [HR].

If a car does not have the condition, the routine will still flag the car with a probability of [FA].

A manager in the company wonders about how to treat the cars that were flagged by the routine, i.e. whether these cars should be offered (at a potentially lower price) or not. He asks you about what to expect about these cars.

What is the probability that any car among those flagged by the mechanism has the production fault?

3.3.8 Scenario 2.3: Factory, severe consequences, statistical. Imagine that there is a factory producing cars that is concerned about a specific fault in the production process. A car with this fault will show no problems until it reaches a mileage of about 3000 miles.

At this point, the production fault will cause affected cars to develop a problem in the steering system. This change might cause cars to block any attempt to change their direction so that severe accidents might follow and the cars cannot be regarded as safe.

The probability to have this production fault in any car coming out of the production line is [BR].

A quality control routine has been implemented to check whether a has the production fault or not. If a car has the fault, the routine will flag this car with a probability of [HR].

If a car does not have the condition, the routine will still flag the car with a probability of [FA].

A manager in the company wonders about how to treat the cars that were flagged by the routine, i.e. whether these cars should be offered (at a potentially lower price) or not. He asks you about what to expect about these cars.

What is the probability that any car among those flagged by the mechanism has the production fault?

3.3.9 Scenario 2.4: Factory, severe consequences, individualized. Imagine that there is a factory producing cars that is concerned about a specific fault in the production process. A car with this fault will show no problems until it reaches a mileage of about 3000 miles.

At this point, the production fault will cause affected cars to develop a problem in the steering system. This change might cause cars to block any attempt to change their direction so that severe accidents might follow and the cars cannot be regarded as safe.

The probability to have this production fault in any car coming out of the production line is [BR].

A quality control routine has been implemented to check whether a has the production fault or not. If a car has the fault, the routine will flag this car with a probability of [HR].

If a car does not have the condition, the routine will still flag the car with a probability of [FA].

Now imagine that you have been offered a car from this factory at a discount price. You obtain the information that the car was flagged by the quality control routine. You wonder what to expect about this car.

What is the probability that this car has the production fault?

3.4 Postquestionnaire

The same postquestionnaire measures were presented as in Study 4a.

4 Details for Study 4c

The estimates for Study 4c will be found on the Harvard Dataverse (https://doi.org/10.7910/DVN/FYMODJ).

4.1 Sample

Note that 1,341 participants started the test, 1,066 passed the initial screening, and 1,016 completed the judgment task. Three responses were removed from the dataset, as the answers mirrored the example given in the instructions that was chosen to be unrelated to the numbers presented in the task, resulting in a final dataset of 1,013 participants. 1,010 participants answered all questions in the surveys.

Table S4
Tasks in Study 4c

Task	b	h	f
1	0.25	0.70	0.40
2	0.04	0.80	0.25
3	0.45	0.85	0.10
4	0.70	0.10	0.25
5	0.85	0.45	0.10
6	0.70	0.60	0.20
7	0.90	0.45	0.35
8	0.95	0.02	0.60

Regarding the 1,013 participants who made an inference, 538 were male (53.1%), 474 female (46.8%) and 1 neither male nor female (0.1%). The age range was between 18 and 81 years, with a mean of 36.7 years ($SD = 11.89$).

4.1.1 MTurk Specification.

- Time estimate: 12–5 minutes
- Fixed payment: $1.50
- Variable additional payment: up to $0.20
- Filters: HIT approval rating $\geq 96\%$, Location: US (also checked via VPS-filter)
- Attention checks: Participants had to pass two out of three consecutive attention checks at the beginning of the study

4.1.2 MTurk Context. In Study 4c, the only task before the Bayes section was another condition of the promising task (Woike & Kanngiesser, 2019) and a demographics section.

4.2 Method

Each participant made a single inference, using the same scenario with randomly chosen probabilities out of a set of eight, listed in Table S4.

Note that the scenario text contains two numbers that are not needed for calculating the solution (the number of months and the corruption rate). These numbers served as distractors in the postquestionnaire.

4.3 Scenario

See Figure S15 for a screenshot of the question.

Imagine that there is a potential mechanical defect in a production line of computers. A computer with this defect will operate normally until it reaches an age of about **8** months.

At this stage, the condition will cause affected computers to go through random freezes and hard drive corruption of up to **98%** of the stored data.

The probability to have this mechanical defect for any computer in the production line is [BR].

A test procedure has been developed to determine whether a computer has or does not have the defect. If a computer has the defect, the test will be positive with a probability of [HR].

If a computer does not have the defect, the test will be positive with a probability of [FA].

Now imagine that you own one of the computers in this production line. The result of the test for your computer came back positive.

What is the probability that your computer has the mechanical defect?

Please enter the probability as a percentage (just the number between 1 and 100, without "%"). Note that an answer of "0.13" would be interpreted as 0.13%=0.0013, whereas "13" is interpreted as 13%=0.13. To prevent misunderstandings we set the mininum for the answer to 1: If you believe that the answer is smaller than 1% (or 0.01), then enter 1.

[*Text entry field.*]

4.4 Postquestionnaire

The postquestionnaire in Study 4c consists of a single question, asking for the consideration of cues in the task. In contrast to Study 4a and 4b, participants were asked to click on the numbers directly, with a zone around each of the numbers responding to mouse overlay and toggling its color after mouse clicks. See Figure S16 for a screenshot of the question. The two distractors were selectable in the same way as the three standard numbers.

Imagine that there is a potential mechanical defect in a production line of computers. A computer with this defect will operate normally until it reaches an age of about **8** months.

At this stage, the condition will cause affected computers to go through random freezes and hard drive corruption of up to **98%** of the stored data.

The probability to have this mechanical defect for any computer in the production line is **25%**.

A test procedure has been developed to determine whether a computer has or does not have the defect. If a computer has the defect, the test will be positive with a probability of **70%**.

If a computer does not have the defect, the test will be positive with a probability of **40%**.

Now imagine that you own one of the computers in this production line. The result of the test for your computer came back positive.

What is the probability that your computer has the mechanical defect?

Please enter the probability as a percentage (just the number between 1* and 100, without "%").
***Note that an answer of "0.13" would be interpreted as 0.13%=0.0013, whereas "13" is interpreted as 13%=0.13.**
To prevent misunderstandings we set the minimum for the answer to 1: if you believe that the answer is smaller than 1% (or 0.01), then enter 1.

Figure S15. Inference in Study 4c

Below is the same task that you responded to before.
Please mark the number or the numbers that you used when giving your response by clicking on the number(s) **before moving on to the next page.** The chosen number(s) will turn green.

Imagine that there is a potential mechanical defect in a production line of computers. A computer with this defect will operate normally until it reaches an age of about **8** months.

At this stage, the condition will cause affected computers to go through random freezes and hard drive corruption of up to **98%** of the stored data.

The probability to have this mechanical defect for any computer in the production line is **25%.**

A test procedure has been developed to determine whether a computer has or does not have the defect. If a computer has the defect, the test will be positive with a probability of **70%.**

If a computer does not have the defect, the test will be positive with a probability of **40%.**

Now imagine that you own one of the computers in this production line. The result of the test for your computer came back positive.

>>

Figure S16. Postquestionnaire Study 4c: Cue use

5 Additional Results for Studies 4a, 4b and 4c

5.1 Judgment classification by task and scenario in Study 4a

Table S5 presents the proportions of responses classified as consistent with the major strategies, split by scenario and probability set (tasks). The distribution demonstrates that probability sets are more important than tasks in determining proportions. This effect goes beyond the trivial finding that multiple strategies' responses can overlap for some probability values (such as the 50% strategy overlapping with the BO in set 1 and HO in set 4). See Table S2 for the probabilities.

Table S5

Classification of respondents in Study 4a split by probability set and scenario. WA_j refers to the WA strategy using Juslin's parameters, WA_m to the WA strategy fitted to all responses in Study 4a.

Probabilities	Scenario	REP	BO	FC	LS	JO	50%	Bayes	WA_j	WA_m
Prob1	Poker	28	26	17	12	12	26	5	0	0
	Dinner	31	25	15	10	10	25	2	0	0
	Car	18	30	18	10	10	30	4	0	0
	Urn	17	48	11	12	12	48	5	0	0
Prob2	Poker	34	25	9	1	31	9	1	0	0
	Dinner	34	20	9	2	26	9	3	0	0
	Car	30	23	16	2	27	16	1	0	0
	Urn	33	29	9	1	31	9	2	0	0
Prob3	Poker	34	36	0	36	2	5	1	2	0
	Dinner	24	38	1	38	7	2	2	2	0
	Car	36	35	1	35	3	1	3	3	0
	Urn	19	47	2	47	4	4	2	2	0
Prob4	Poker	33	34	9	4	4	33	9	0	0
	Dinner	31	27	9	9	5	31	9	0	0
	Car	29	29	4	9	7	29	4	0	0
	Urn	22	25	13	6	13	22	13	0	0

5.2 Classification and demographics

Table S6 gives the proportion of female participants and participants with a postgraduate degree or PhD, split by the classification of judgments in Studies 4a and 4b combined. Participants were classified, when one and only one of the listed strategies was consistent with at least 50% of their responses, they were categorized as "unclassified", otherwise.

Table S6
Classification, education, and gender (sorting criterion) in Studies 4a and 4b

Classification	Female [%]	Postgraduate [%]	N
JO	33	25	24
Bayes	37	37	30
FC	37	16	38
Uncl.	49	12	331
50%	56	26	27
LS	58	21	24
REP	63	13	186
BO	64	16	191
f	75	8	12
Total	55	15	863

5.3 Classification and self-classification

Table S7 summarizes the correspondence of objective and subjective categorizations of strategy use. For the most common strategies, the majority of participants were able to pick out their own strategy when presented with its (formal) representation. This result is striking, in that it rules out the possibility that responses are predominantly due to confusion or random responding. For example, 66% of participants who respond with the base rate, indicate that they actually chose the base rate as response, when the problem was presented in abstract terms.

Table S7
Classification based on judgments (rows) and self-classification (columns) by participants: Numbers present rounded percentages of matching classifications

	Self-classification (multiple possible)									
Classification	Bayes	JO	LS	HO	BO	FC	FO	no system	WA	N
Bayes	**67**	3	3	3	0	0	0	0	7	30
JO	0	**79**	8	13	8	4	0	4	13	24
LS	0	0	**57**	17	17	0	0	13	0	23
REP	2	8	5	**53**	27	1	10	13	2	182
BO	1	6	6	18	**66**	1	3	12	1	190
FC	3	8	24	13	11	**29**	3	13	8	38
FO	8	8	8	25	17	8	**33**	25	0	12
50%	0	7	4	19	33	0	4	**33**	0	27
Unc	5	9	13	27	33	2	10	**17**	6	321

5.4 Postquestionnaire

5.4.1 Statistical self-evaluation and comparison. Table S8 reports answers to the self-assessment questions, conditional on the classification of answers. For example, participants whose responses were classified as corresponding to Bayes' theorem had predominantly taken a statistics class (83%), liked tasks involving calculations (90%) and correctly considered their performance to be above average (93%). Seventeen participants made all judgments in the judgment task but did not complete the postquestionnaire.

Table S8

Responses to postquestionnaire questions in Studies 4a and 4b conditional on classification: Columns report the rounded percentages of respondents in each classification category for (1) having attended a statistics class (rows are sorted by this percentage), (2) being able to apply Bayes' theorem , (3) liking calculations (two affirmative categories), and (4) believing to be in the top half of performers

Classification	Statistics class [%]	Can apply theorem [%]	Likes calculations [%]	Above average (belief) [%]	N
Bayes	83	27	90	93	30
JO	63	4	83	54	24
FC	58	5	68	63	38
LS	48	0	39	54	23
f	42	0	42	75	12
Uncl.	40	3	60	45	320
REP	33	3	49	53	182
BO	33	2	56	57	190
50%	26	7	67	48	27
Total	40	3	58	53	846

5.4.2 Open-format answers (examples). Here, we list some of the responses given to open questions, when commenting on the HIT or when detailing systematic or unsystematic strategies. As many participants did not respond to these questions, we present only selective examples without estimating relative frequencies or making claims of representativeness.

Many of the open-format explanations matched specific strategies, such as the following:

- I chose the Base Rate exactly.
- I figured that the base rate is the actual chance of it being a positive result, since none of the test results would be 100%, so it was best to just stick to that number.
- To me the questions were asking for the base rate because of the way the questions were worded.
- Yes. As I mentioned above, I think the Hit rate illustrates the probability. The parameters were straightforward and I feel that I was able to interpret them correctly.
- I chose the hit rate
- I subtracted the FA from 1.0.
- (50% chance it could go either way) .5

Some participants made an effort to describe complex approaches, some but not all of them (see the confusion of $1-b$ with $1-h$ in the second example) resulting in correct judgments.

- I assumed a total population of either 100 or 1000 to make my computations easy. I then created an empty table with two rows and two columns, with rows corresponding to the populations with/without the attribute and columns corresponding to positive/negaive test results. I then multipled the total population by the base rate to compute the number of population members with the specific attribute; I subtracted that number from the total population to compute the number of population members without the atribute. Using the hit rate, I multiplied to compute the number of members of the population with the attribute and a positive test result, then subtracted that from the total number of members of the population with the attribute to get the number of membrs with the attribute and a negative test result. I entered those values in the top row of my table. Similarly, using the false alarm rate, I computed the numbers of members without the attribute and positive/negative test results and [...].

- I imagined a sample size of 10000. Then I multiplied 10000 by the br to find how many people have the disease. Then I took that number and multiplied by the hit rate to find the number of people with the disease who test positive. Then I took the numbe of people who don't have the disease (10000 - 10000 X HR) and multiplied by the FA. This give the number of people who would test positive if even if they do not have the disease. Then I added these two numbers together to get the total number of peopl who test positive and I calculate the ratio of the number who tested positive and have the disease to the total number of people who tested positive.

- I used Bayes Rule to calculate these probabilities! I recognized these problems right away, just had to refresh my memory on Bayes Rule.

Other participants demonstrated changes in their strategies or described how they deviated from the results of calculations.

- I kinda used either the BR and HR or HR and FA or all three, depending on how relevant I thought they were.

- I used (HR + FA)/2 but used common sense when evaluating

- Kind of HR-FA, but I knew that wasn't right, so I tweaked it a little

- For the first 2 I combined and averaged the HR and FA, I later realized this couldn't be correct and stuck with the BR as the overall answer.

- I calculated HR-FA if BR was irrelvant and (HR-FA) x BR when BR was relevant, BR being relevant if it was not known in the specific example.

- At first I tried to find the average between two numbers usually the BR and HR. Later I simply chose the BR or HR based onw which I felt was more relevant.
- On the last scenario I added the HR plus the FA then multiplied the total by the BR. (HR+FA)xBR
- unfortunately, I changed strategies across scenarios, ending up with BR x HR
- The first one I meant to chose the BR but I think chose the HR by mistake. I realized that when I did the 2nd task. When I was in the previous screen I forgot to mention that. I think that the other rates do not affect the original odds and are red herrigs.

A smaller group of participants indicated that they did not feel competent to solve the problems or did not want to invest the time to solve them appropriately.

- No. I didn't know the correct way to solve the problems, so I just guessed.
- I have no interest in doing this kind of math, since it's really not accurate and real life has much more complexity to it to bother with such overly detailed approaches. The detailed stuff is best left to engineers trying to build stuff, and I'm no engieer. :-) I like to stick to -1, 0, and 1, with maybe a 2 or 3 thrown in if absolutely necessary!
- I played with it a bit including all 3 factors, but I didn't have the patience at the moment to really dig in. and compute. I'm usually pretty logical, but today I just didn't feel the challenge. I took a stats class, but it wasn't my strongest one by a ong shot.
- I think I'm mostly just not very good with probability. On the first one, the fact that the meal was enjoyed with probability of 0.50, it seemed I coud just eliminate that and find use the other data. On the other two it seemed like the HR and FA were irrlevant. It seemed like the question basically read, "The probability of x = y. What is the probability of x?" so I answerd "y".
- I am not good at math and have never studied probability.

5.5 Response times in Studies 4a, 4b, and 4c

Figure S17 shows that the time needed to respond systematically varied across rules. Those whose inferences were predicted by Bayes' rule clearly took longest.

Differences between rules requiring two pieces of information (joint occurrence and likelihood subtraction) and rules requiring a single piece of information (false-alram complement, representativeness and baserate only) were less clear, and difficult to generalize based on the small sample sizes. Rules that require the mere repetition of given information (REP and BO) tended to be associated with faster response times (but note the fast responses for JO in Study 4b based on two participants). It should also be considered that the measurement of response times

included the time spent on reading and comprehending the task and participants self-selected into groups, making it unlikely to see a perfect pattern (Study 4c with one task and the largest sample size comes closest to exhibit an ideal differentiation between the three groups). In conclusion, both analyses of self-reported information use and response times lend support to the classification: Bayes' rule with three cues is clearly associated with the longest response times.

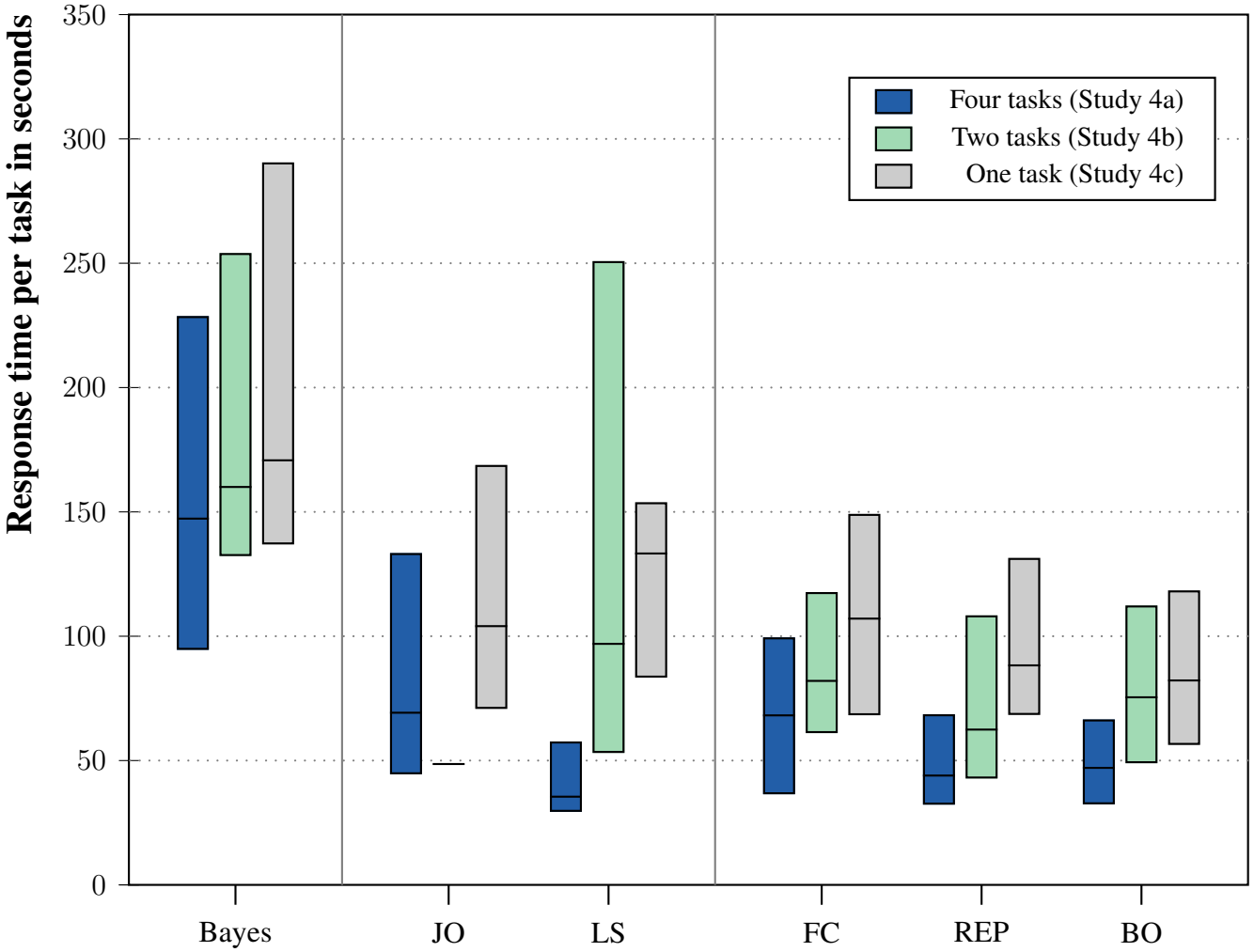

Figure S17. Medians and quartiles for response times per task for each category of participants in Studies 4a, 4b, and 4c. Rules are presented on the x-axis; bars mark the interquartile distance, with horizontal lines showing the median response times.

6 Additional Results for Study 5

6.1 Illustration of error distributions

We used two different methods of creating estimate distributions for single-process models. In both cases, errors are normally distributed with mean 0 and standard deviations s, applied

to original estimates in two different ways.

Simple errors (exemplified in Figure S18) are directly added to probability estimates. Results above 1 are set to 1 and results below 0 are set to 0. In a second step, we determine the proportion of estimates for each of the 101 intervals. Note that, as before, the 99 interior intervals cover a range of 0.01, the two outer intervals a range of 0.005. The middle interval is $[0.495; 0.505[$. The figure demonstrates estimate distributions for two different error-free estimates ($O = 0.1$ and $O = 0.7$) and six different standard deviations. With a standard deviation of 0, the model predicts that all observations fall into a single interval (the y-axis is limited to 0.5 to benefit the visualization of the remaining scenarios). With an increase in the standard deviation, the predictions spans more and more intervals. For $M \neq 0.5$, the predictions are not symmetric around the error-free value, as values larger than 1 and lower than 0 are considered to fall into the outer intervals (explaining the peaks in these intervals observed for high standard deviations). The mean prediction is therefore also not generally equivalent to the error-free prediction (O).

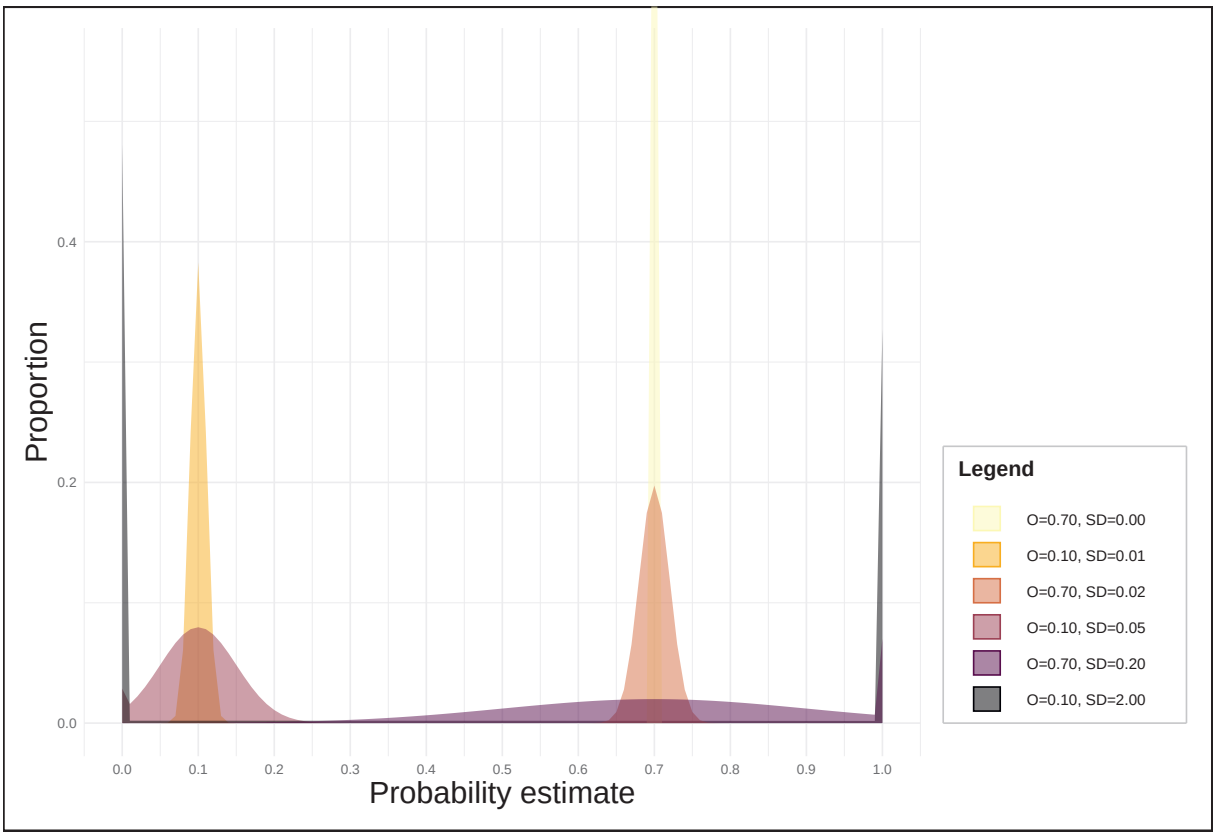

Figure S18. Examples for simple error distribution: The upper lines of colored areas connect the 101 interval estimates for six parameter combinations. Estimates are normally distributed with mean O and standard deviation SD, but estimates above 1 and below 0 are set to 1 and 0, respectively.

Log-odds errors are applied to the log odds of error-free estimates ($ln(\frac{p}{1-p})$). The resulting log-odds are then retransformed into probabilities. Predictions are then summed up for each interval. In contrast to the first method, the resulting probabilities are guaranteed to be within the interval $[0; 1]$. At the same time, we still consider responses of 0 and 1 to be legitimate matches for the two extreme intervals. The non-linear transformation introduces skewness: Deviations towards the center ($p = 0.5$) are more likely than the same deviations towards the respective end. Some possible distributions of predictions are demonstrated in Figure S19, using the same two original values as for the simple error distributions. Standard deviations need to be higher to achieve errors of similar size.

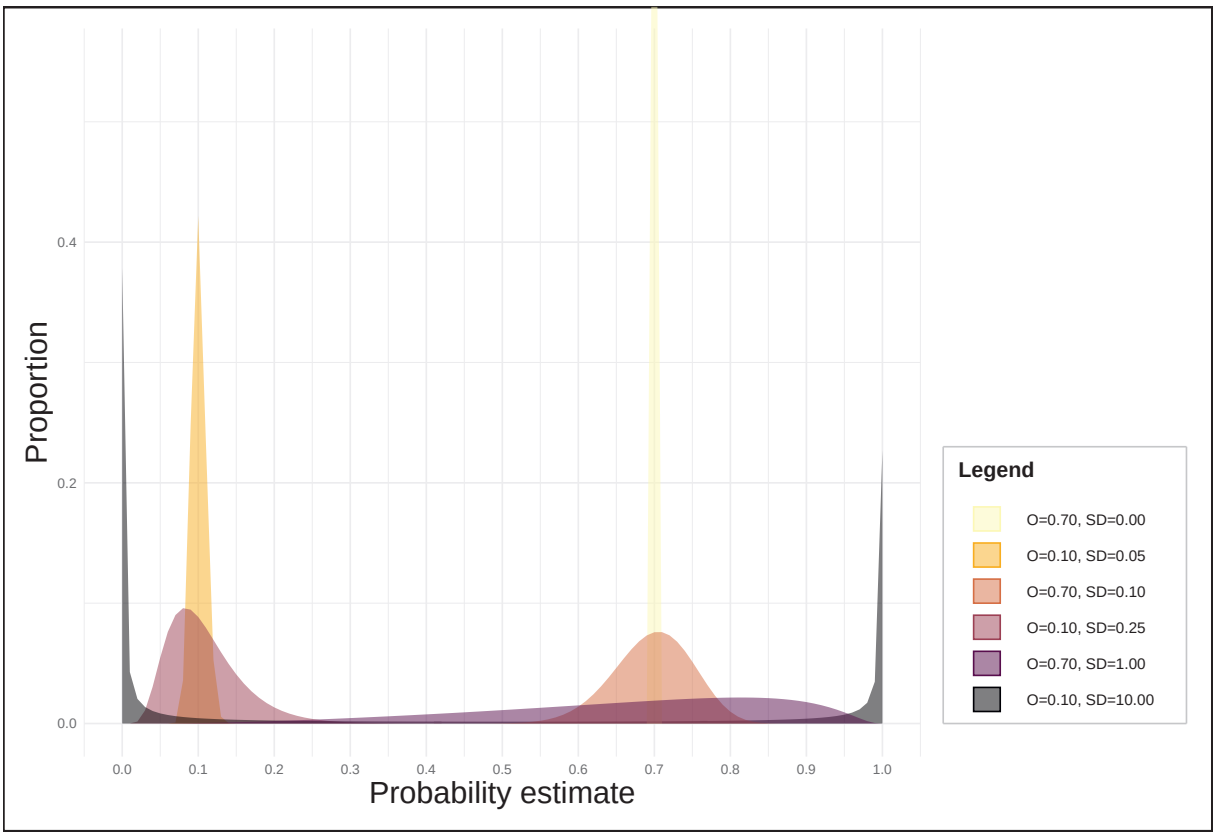

Figure S19. Examples for log-odds error distribution: The upper lines of colored areas connect the 101 interval estimates for six parameter combinations. Errors are normally distributed with mean 0 and standard deviation SD and added to the log-odds of probability O, results are retransformed into probabilities.

6.2 Overlap optimization procedure

The transformations and calculations involved in determining the overlap criterion does not allow for a fast and unambiguous single procedure. We therefore used a sequence of steps

for optimization that differed slightly depending on the number of parameters. Models had either one or two free parameters to be chosen in this way, and the procedure varied slightly between these cases.

6.2.1 One-parameter optimization. Most models had one parameter, namely the standard deviation of the error distribution. In these cases, we went through six steps.

1. We used fmincon, a non-linear optimization function in Matlab. The boundaries for the search interval were set to 0 and 1, and 0.1 was chosen as starting value for simple errors, an interval of 0 to 5 with 0.2 as starting value for log-odds errors.

2. We repeated this procedure with patternsearch, another Matlab optimization routine.

3. We determined the best performance for all parameter values in a grid search, moving from 0 to 1 in steps of 0.001.

4. We searched the vicinity of the previously determined solution with a finer grid, investigating all parameter values within one step size from the previously found solution (above or below) in steps of 0.0001.

5. We used the result of the previous step as starting value for fmincon.

6. We used the same result as starting value for patternsearch.

As the purpose of this task was optimization, we then determined the best solution across all of these steps (the fourth step is guaranteed to achieve at least the performance of the third). For some of the postdiction models, we added steps in the case that the parameter was estimated at the upper boundary (1 or 5). In these cases we repeated the procedure with a coarser grid in steps 3 and 4 for the interval 0 to 20 for simple errors, and 0 to 50 for log-odds errors. In general, results confirmed the usefulness of multiple variants, as different solutions were chose across rules and tasks. Differences between best and second-best solutions were reassuringly small.

6.2.2 Two-parameter optimization. Some conservatism model optimized for joint values of two parameters, adding the relative weight of the base rate (vs. the Bayesian response) as a second parameter. Likewise, the postdiction models that estimated an optimal single prediction, replaced the conservatism weight with a prediction parameter in the same interval. In these cases, we went through eight steps.

1. We used fmincon with two variables. For the standard deviation parameter, the boundaries for the search interval were again set to 0 and 1, and 0.1 was chosen as starting value for simple errors, an interval of 0 to 5 with 0.2 as starting value for log-odds errors. For the conservatism/prediction parameter, the search interval was [0;1] in all cases, with a starting value of 0.5.

2. We repeated this procedure with patternsearch.

3. We determined the best performance for all parameter values in a grid search. Both parameters were incremented independently in steps of 0.01 (the larger step size was necessary to keep calculation times feasible), the best combination was chosen.

4. Keeping the previous results for the error parameter constant, we varied the conservatism parameter in steps of 0.0001 through the search interval.

5. Keeping the step 3 result for the conservatism/prediction parameter constant, we varied the error parameter in steps of 0.0001 through the search interval.

6. Starting with the step 3 solution, we explored the vicinity of the solution in both dimensions simultaneously, varying the conservatism parameter within the interval of 0.02 above and below the found solution in steps of 0.0005, the error parameter within 0.1 above and below the found solution in steps of 0.001.

7. We used the result of step 6 as starting value for fmincon.

8. We used the results of step 6 as starting value for patternsearch.

Again, we used an extension of the search interval for the error parameter, when these steps estimated the parameter at the upper boundary. Results showed that all of these steps improved the overall performance. The increase in running times discouraged us from exploring models with more than two free parameters to optimize (beyond those parameters minimizing least squares in regression models). The current simulations ran on up to five desktop computers/calculation servers in parallel over the course of weeks.

6.3 Comparison of overlap comparisons within model categories

In the main manuscript, we reported the performance of the best-performing models in each category. Here, we add the comparison of all models by category.

6.3.1 Weighing-and adding and optimal single-process models. First, we show a comparison of weighing-and-adding models with optimal single process models and two reference models (see Figure S20). The two prediction models, WAp amd WAo, only differ in the type of added error. Their performance is very close: the WAo model is slightly better with an average overlap of 18.7% compared to the WAp performance at 18.0%. Similarly, for the two optimal single process models, OSo with an average overlap of 36.7% slightly outperforms OSp 35.7% by one percentage point. The MID model predicts the center of the estimation scale (0.5) with added standard errors estimated to maximize overlap. Its performance at 22.4% falls between the prediction and postdiction models. This demonstrates that the optimal single process models could not reach their level of performance without identifying a good center for the distribution, whereas the weighing-and-adding model do not outperform a model picking the center. The MaxInt model picks the single interval with the largest proportion of cases to contain 100% of the expected estimates. This model with a performance of 29.4% comes within

about 6 percentage points of the the optimal process models. This demonstrates that picking the right center is the most important component in explaining the performance of single-process models.

6.3.2 Conservatism models. The comparison of conservatism models shows a similar grouping as in the previous comparison (see Figure S21). We created four prediction and two postdiction models. The worst performing model was CNf (10.3%), based on a restricted regression with the full dataset. The two models based on finding the optimal weight in the training sample performed better. In this case, standard errors in the CNp model were a better choice (17.1%) than log-odds errors (15.6%). The best prediction model was the regression-based model based on filtered datasets with estimates within the interval between base rate and correct response (18.2%). It should be noted that both CNo and CNp estimate weights close to 1 across all tasks, with a small standard deviation. In this sense, the models devolve to predicting the base rate, and their performance is worse than that of the simple BO model in the section below.

The postdiction models showed an expectedly better performance with an advantage for OCo (31.8%) over OCp (30.3%). Both models are very close to the performance of the MaxInt model in the previous comparison. For the OCp model, the weight parameter is estimated to be $w = 1$ in 17 tasks, and $w = 0$ in 34 tasks.

6.3.3 Representativeness and base-rate models. Six additional models were based on the two best-performing single rules: base-rate only and representativeness, four prediction and two postdiction models (see Figure S22). Simply predicting the base rate (BO) or the hit rate (REP) achieves a performance of 16.0% and 24.2%, respectively, the latter being better than all prediction models considered so far. Adding a simple error component does not improves this performance only slightly with 17.1% for BOp and 24.8% for RPp. The postdiction models achieve a performance of 27.2 for OBO and 30.0% for ORP, close to the performance of OCp and not far behind OSo. In this sense, respresentativeness and conservatism seem to be nearly equally successful, but compared to the performance of the toolbox model, the extent of this success sappears rather limited.

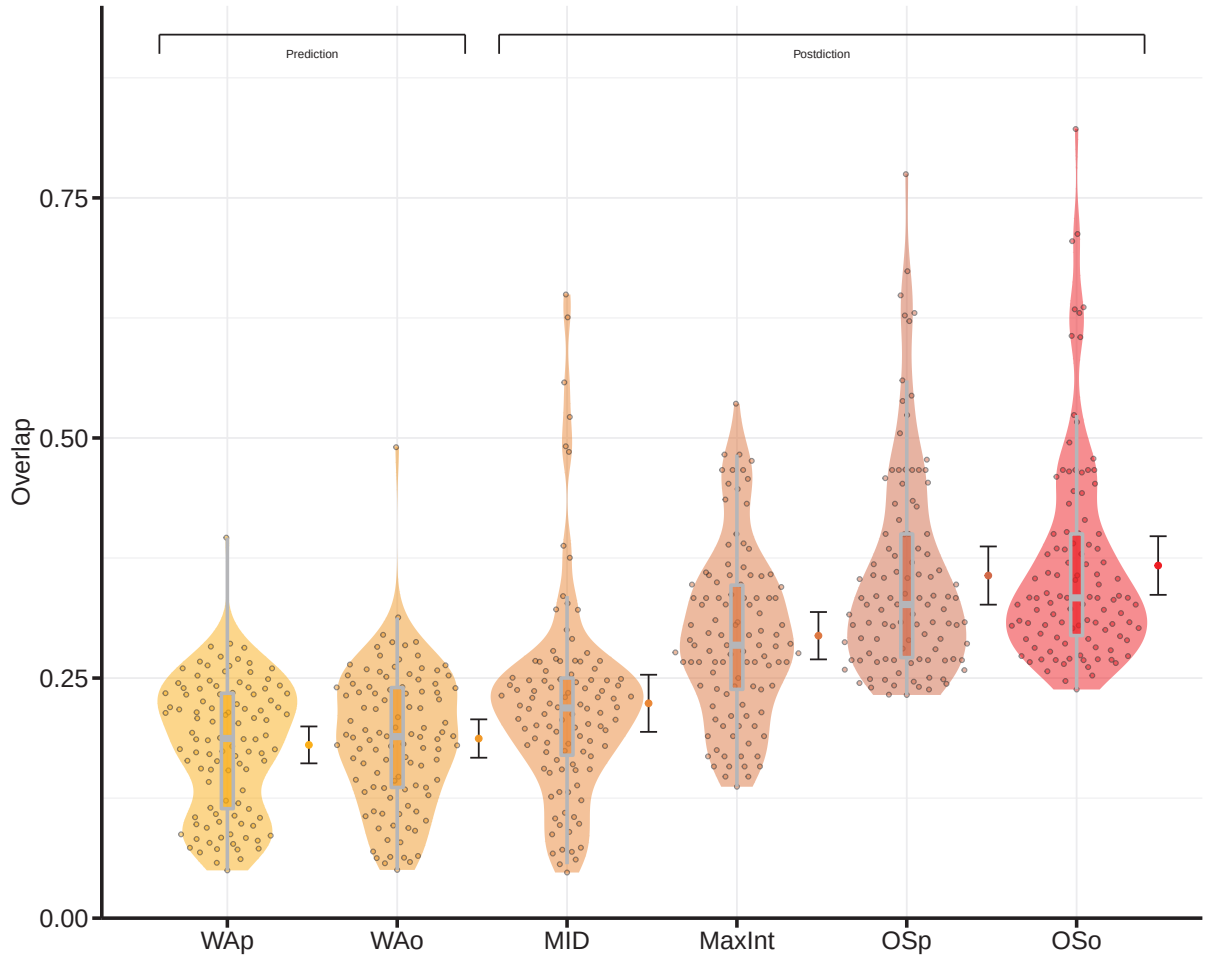

Figure S20. Achieved overlap of selected prediction and postdiction models across datasets: For each model, violin plots, beeswarm plots. and boxplots show the distribution of overlap values across the 106 tasks. Dots and bars to the right of each cloud show means and the 95% CI. Data is presented for WAp: the weighing-and-adding model with added simple errors, WAo: the weighing-and-adding model with log-odds errors, MID: the model predicting 0.5 with simple errors, MaxInt: the model predicting the interval with the most observations, OSp: the optimal single-process prediction with simple errors, and OSo: the optimal single-process prediction with log-odds errors.

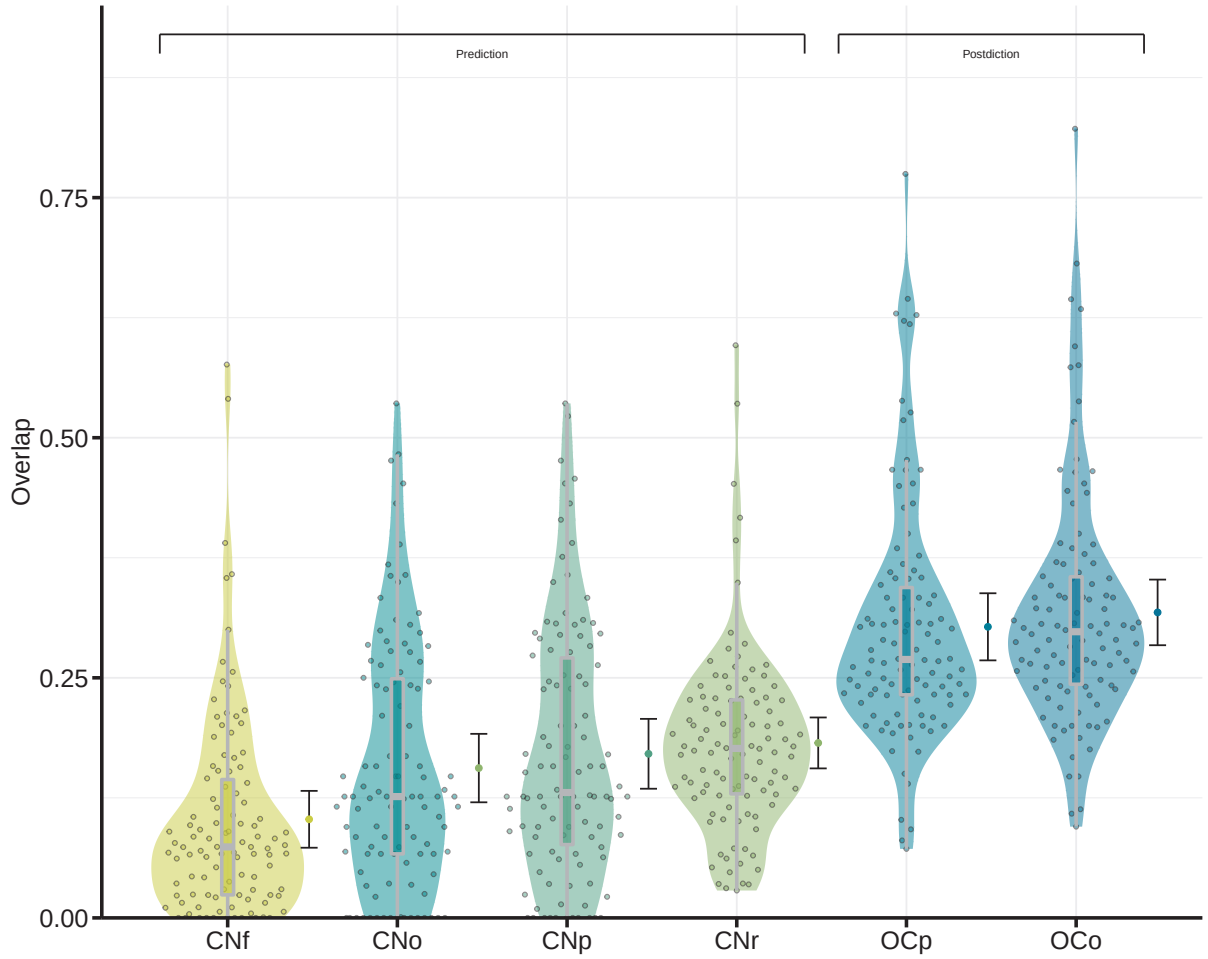

Figure S21. Achieved overlap of selected prediction and postdiction models across datasets: For each model, violin plots, beeswarm plots. and boxplots show the distribution of overlap values across the 106 tasks. Dots and bars to the right of each cloud show means and the 95% CI. Data is presented for CNf: the conservatism model based on full datasets with simple errors, CNo: the conservatism model picking optimal weights in the training sample with log-odds errors, CNo: the conservatism model picking optimal weights in the training sample with simple errors, CNr: the conservatism model based on restricted datasets with simple errors, OCp: the optimal conservatism model with simple errors, and OCo: the optimal conservatism model with log-odds errors.

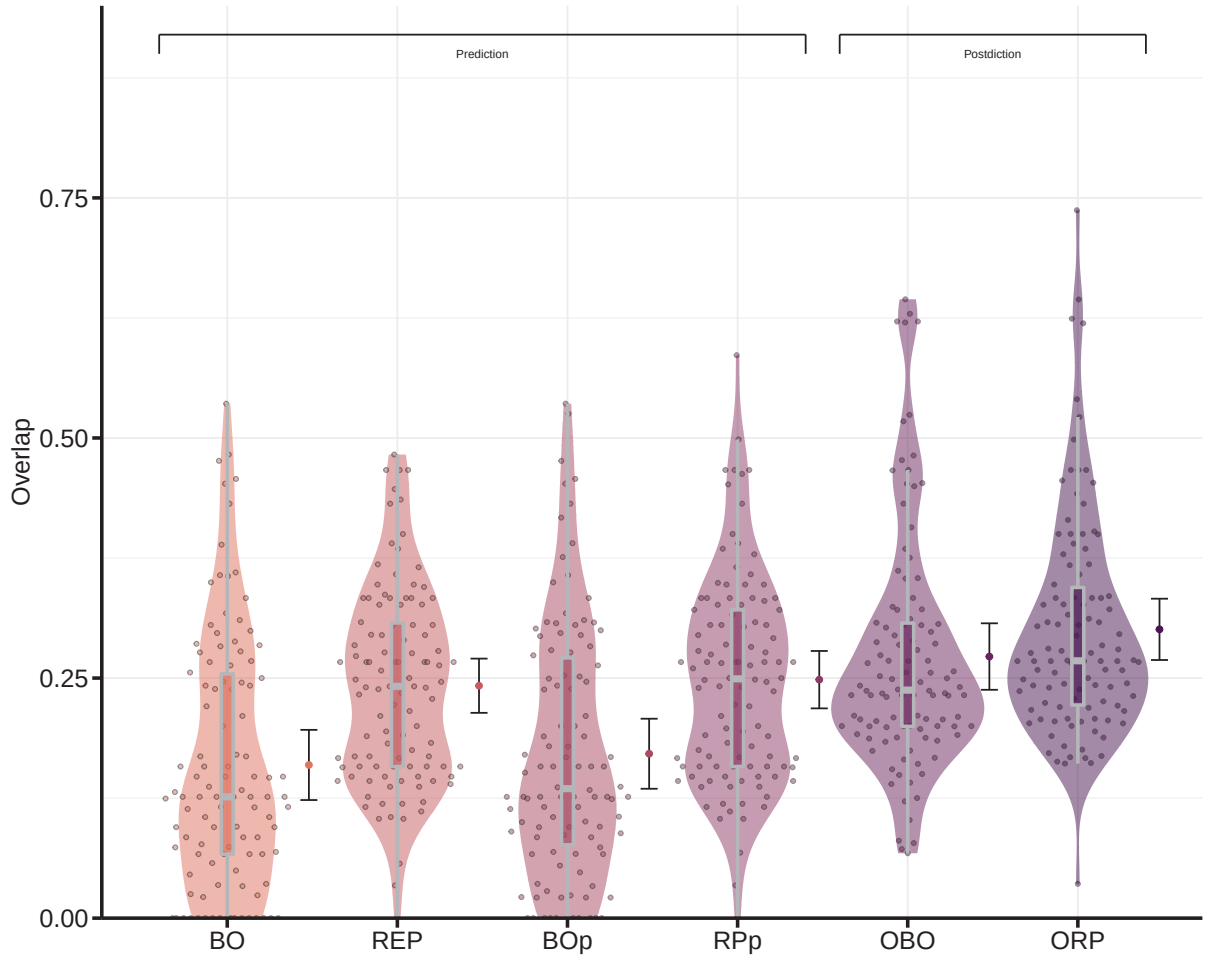

Figure S22. Achieved overlap of selected prediction and postdiction models across datasets: For each model, violin plots, beeswarm plots. and boxplots show the distribution of overlap values across the 106 tasks. Dots and bars to the right of each cloud show means and the 95% CI. Data is presented for BO: predicting the interval containing the base rate, REP: predicting the interval containing the hit rate, BOp: the model predicting he base rate with simple errors, RPp: the model predicting the hit rate with simple errors, OBO: the optimal model predicting the base rate with simple errors, and ORP: the optimal model predicting the hit rate with simple errors.

6.4 Predictions for each task with information dashboards

This section presents dashboards for every single task in Study 5, comparing the toolbox predictions and performance with those of single-process and weighing-and-adding models in the first subsection, and with those of conservatism models in the second subsection. See the main manuscript for an explanation of the dashboard elements.

6.4.1 Comparison of single-process and toolbox predictions. Each plot shows a comparison between estimates and overlap of single-process models and toolbox models for one task (four plots are summarized in one figure). The plots present the distribution of relative response frequencies across intervals, as well as the estimates of the optimal toolbox model (OTB, blue bars) and estimates of the predictive five-plus toolbox model (FPT, dashed lines with whiskers). Overlaid are estimates of the predictive weighing-and-adding model with simple errors (WAp) and the optimal single-process model with log-odds errors (OSo). The dashboards in the upper right summarize task information, the overlap performance of each model, and the overlap performance of the OSo model across intervals not estimated by the OTB.

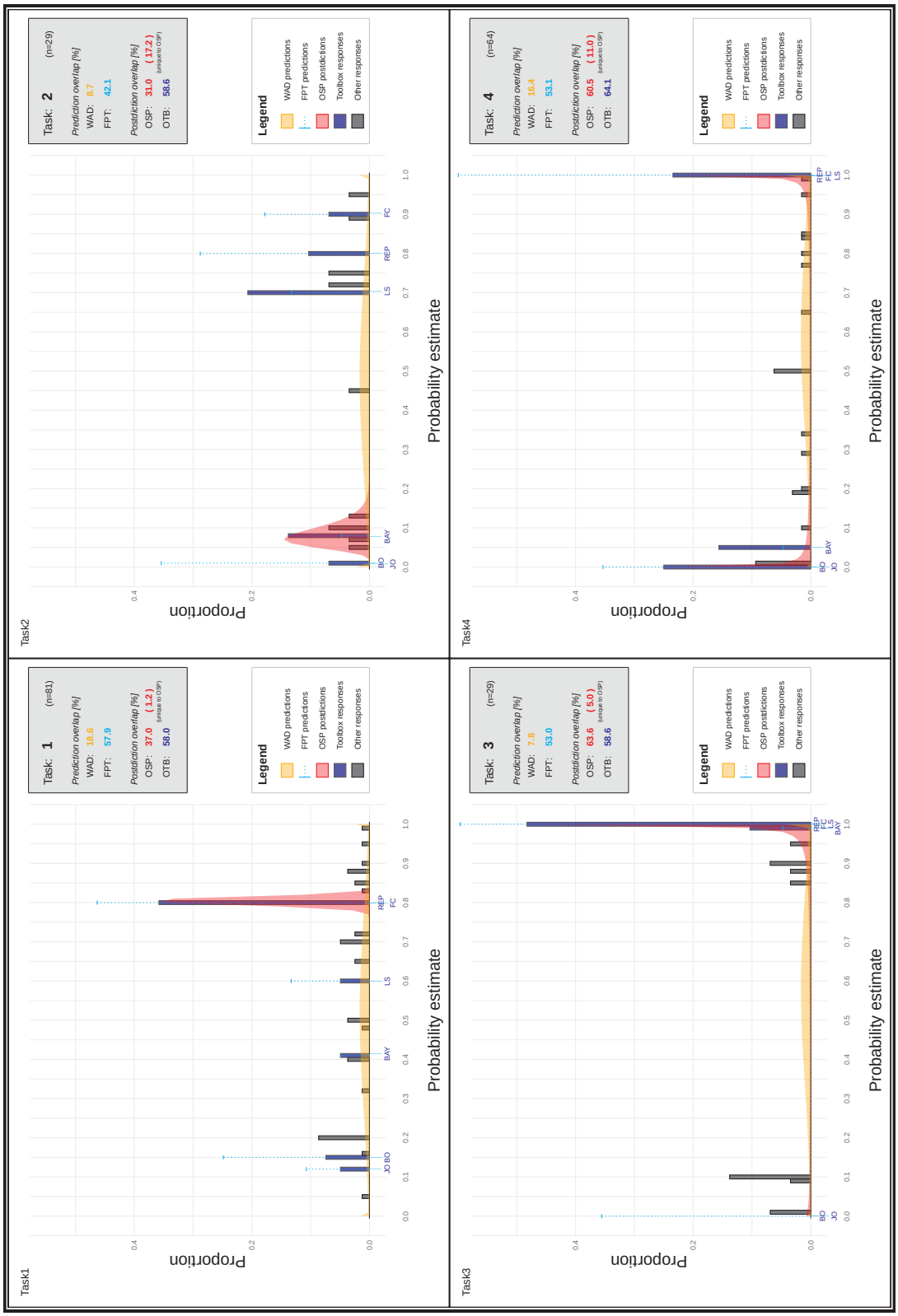

Figure S23. Comparison of toolbox and single-process predictions across tasks (part 1/27)

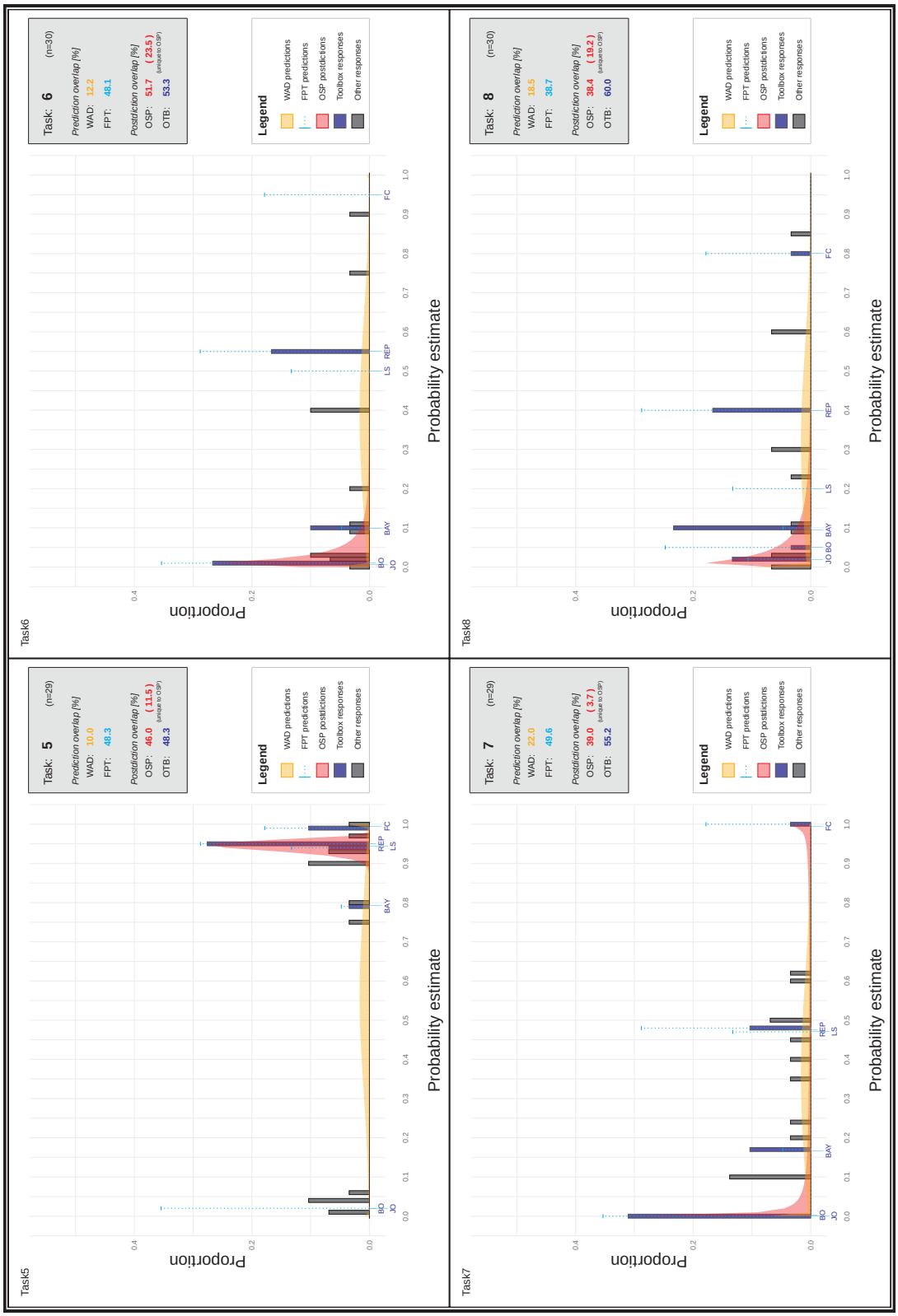

Figure S24. Comparison of toolbox and single-process predictions across tasks (part 2/27)

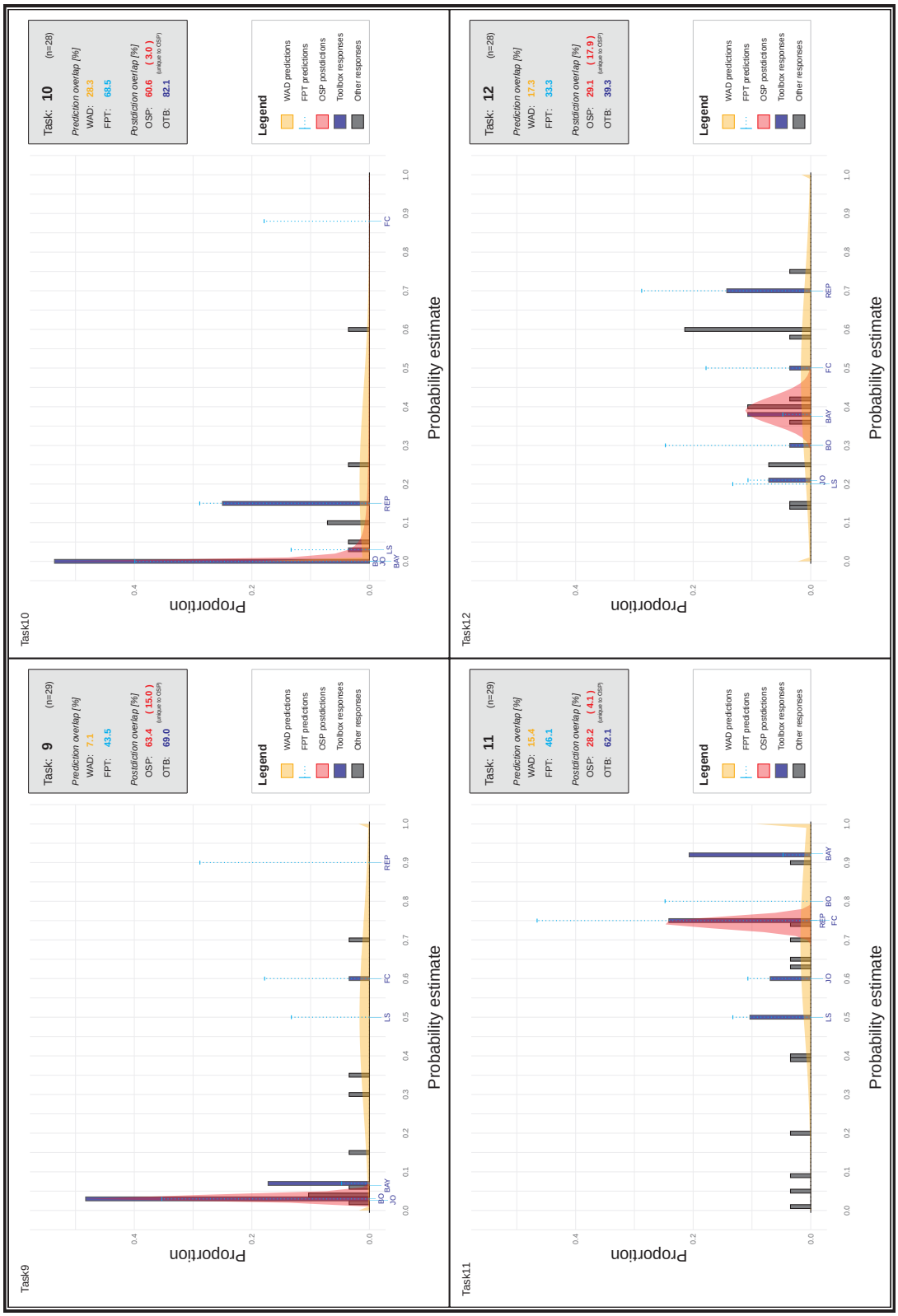

Figure S25. Comparison of toolbox and single-process predictions across tasks (part 3/27)

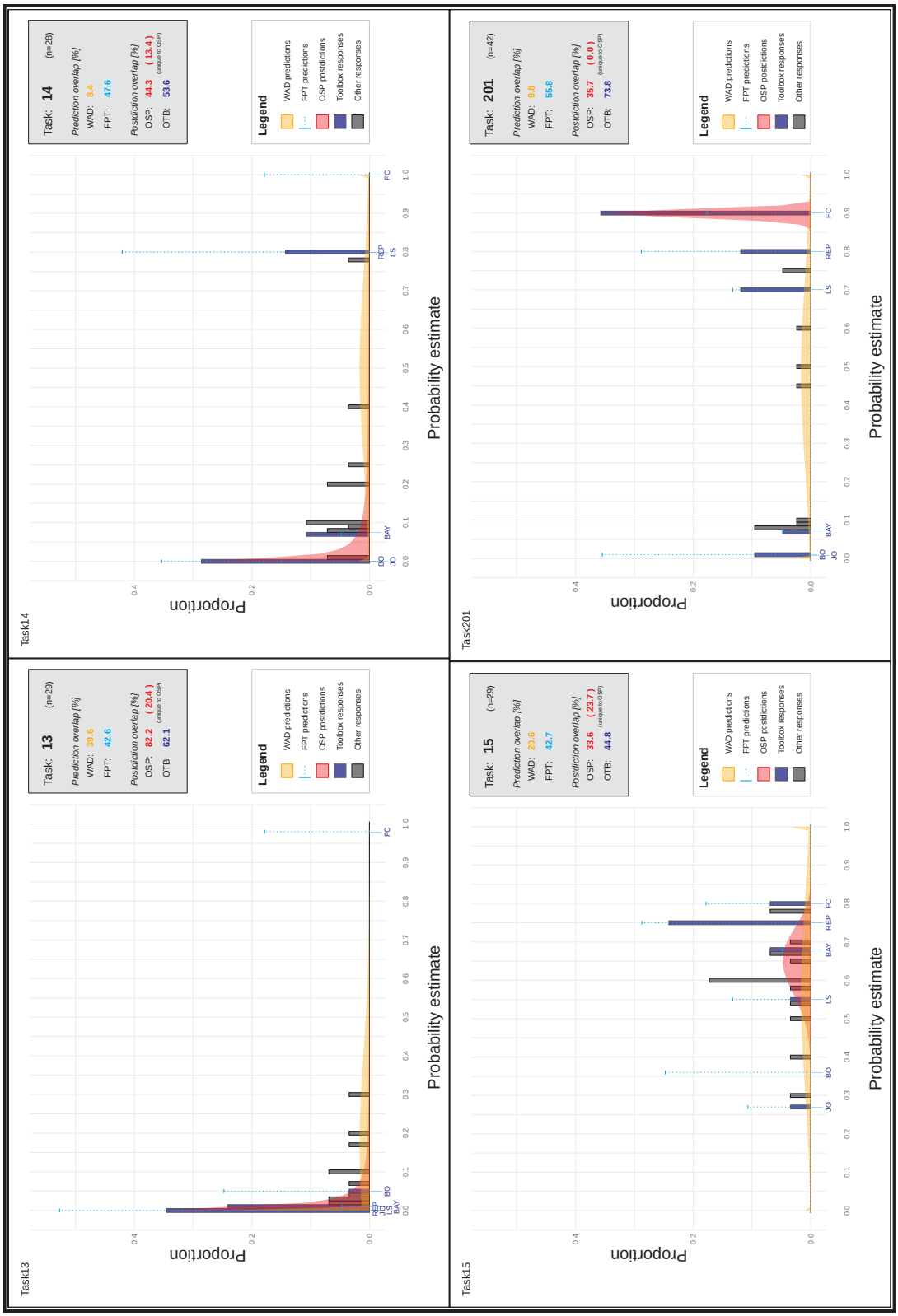

Figure S26. Comparison of toolbox and single-process predictions across tasks (part 4/27)

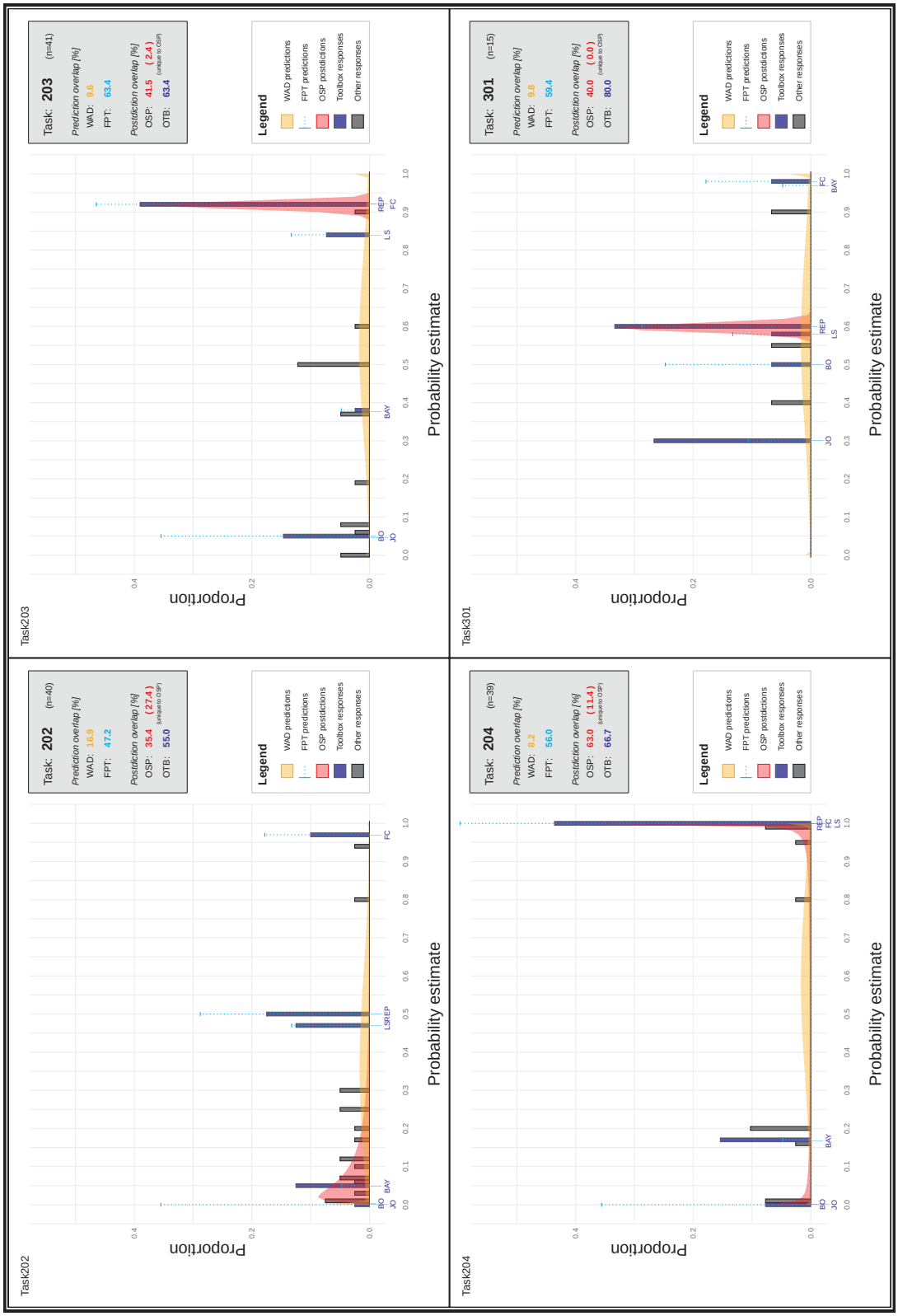

Figure S27. Comparison of toolbox and single-process predictions across tasks (part 5/27)

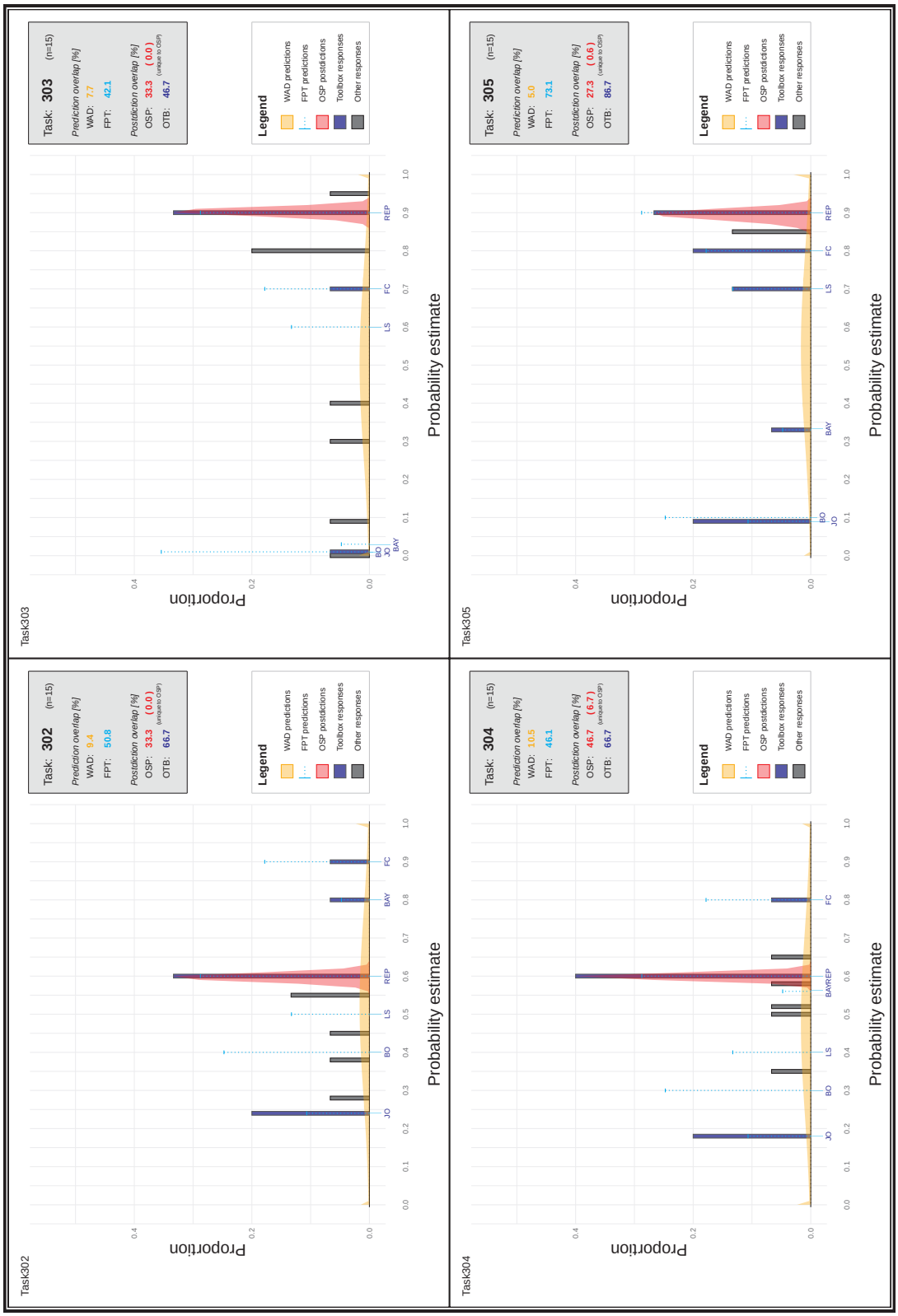

Figure S28. Comparison of toolbox and single-process predictions across tasks (part 6/27)

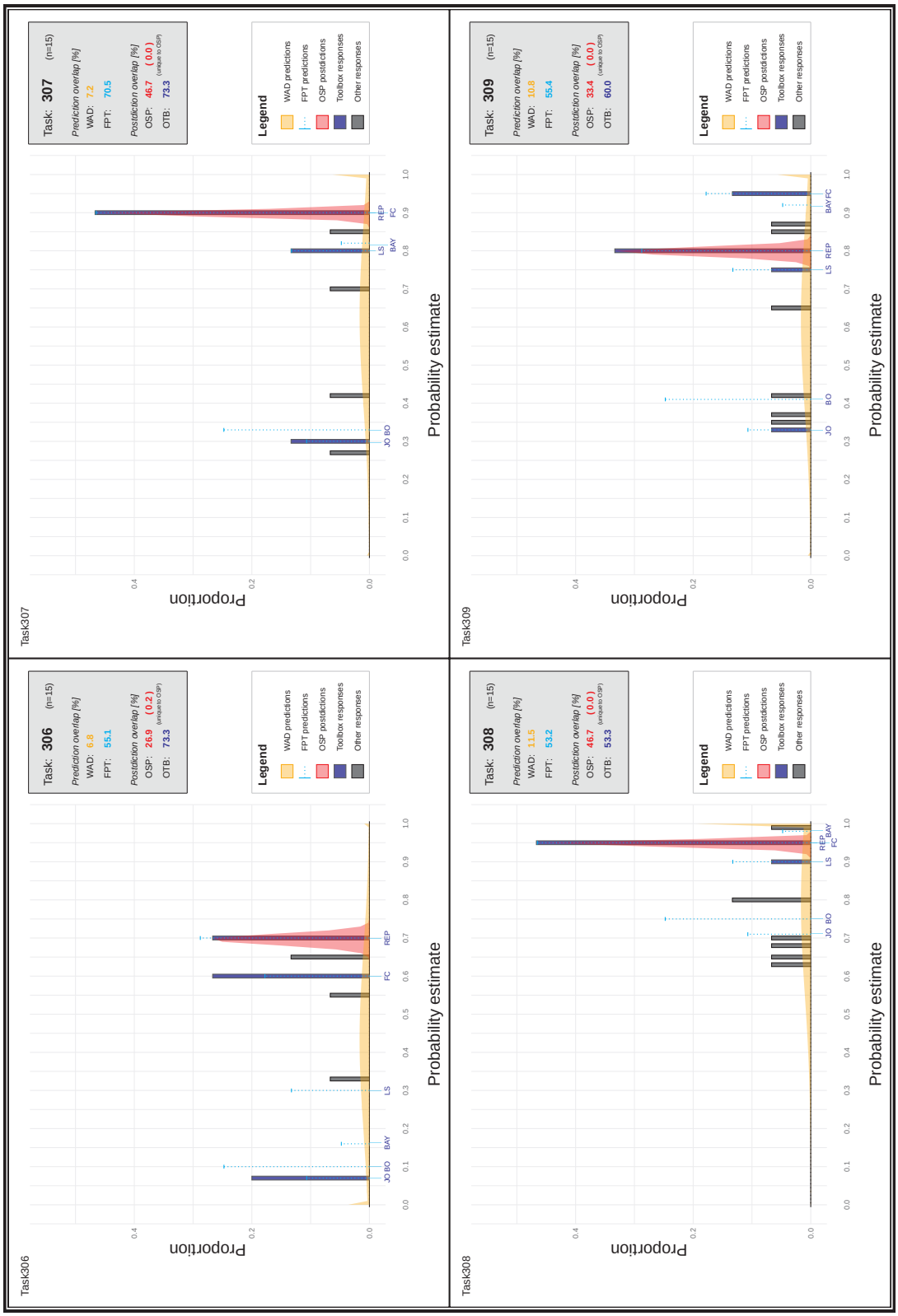

Figure S29. Comparison of toolbox and single-process predictions across tasks (part 7/27)

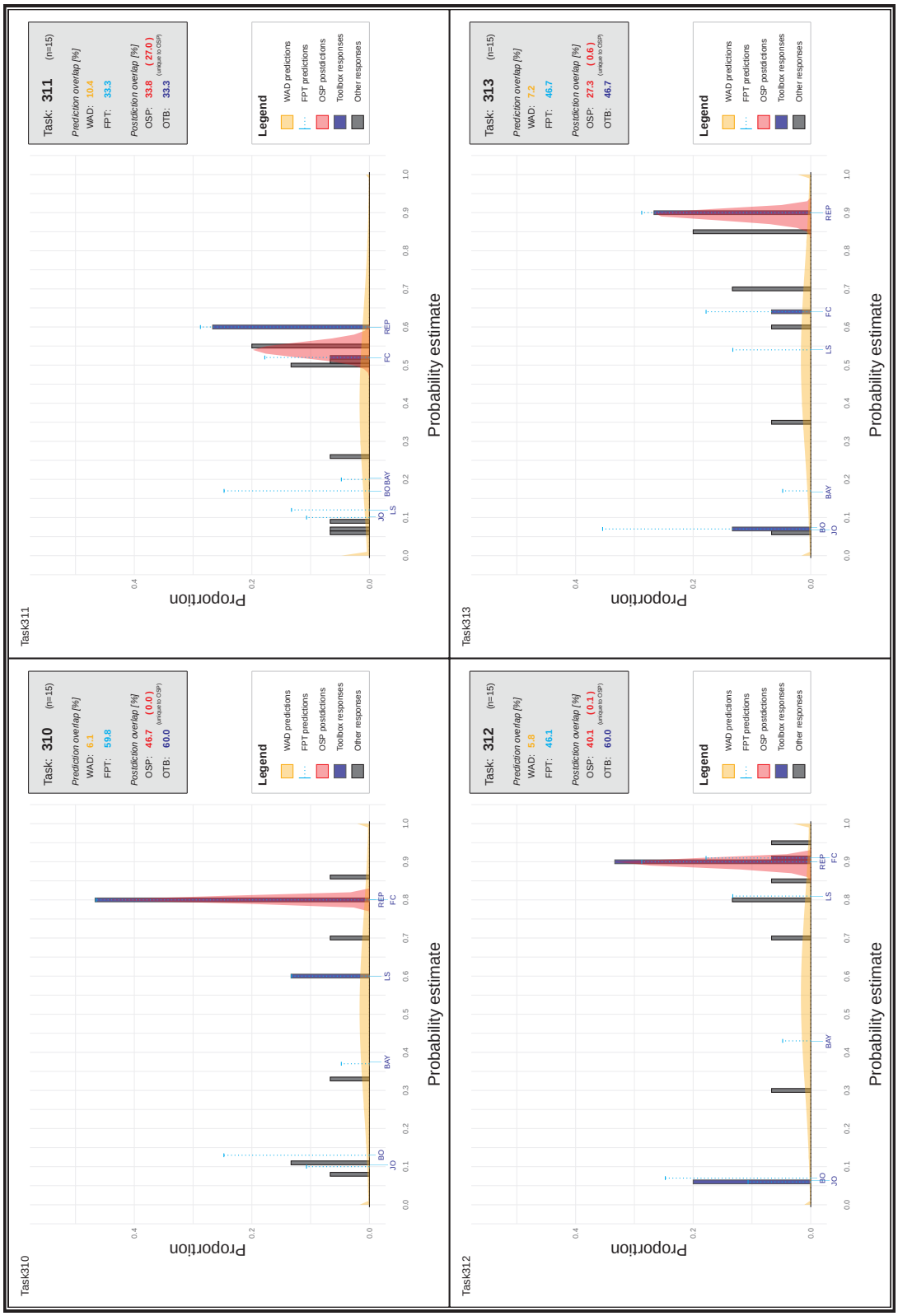

Figure S30. Comparison of toolbox and single-process predictions across tasks (part 8/27)

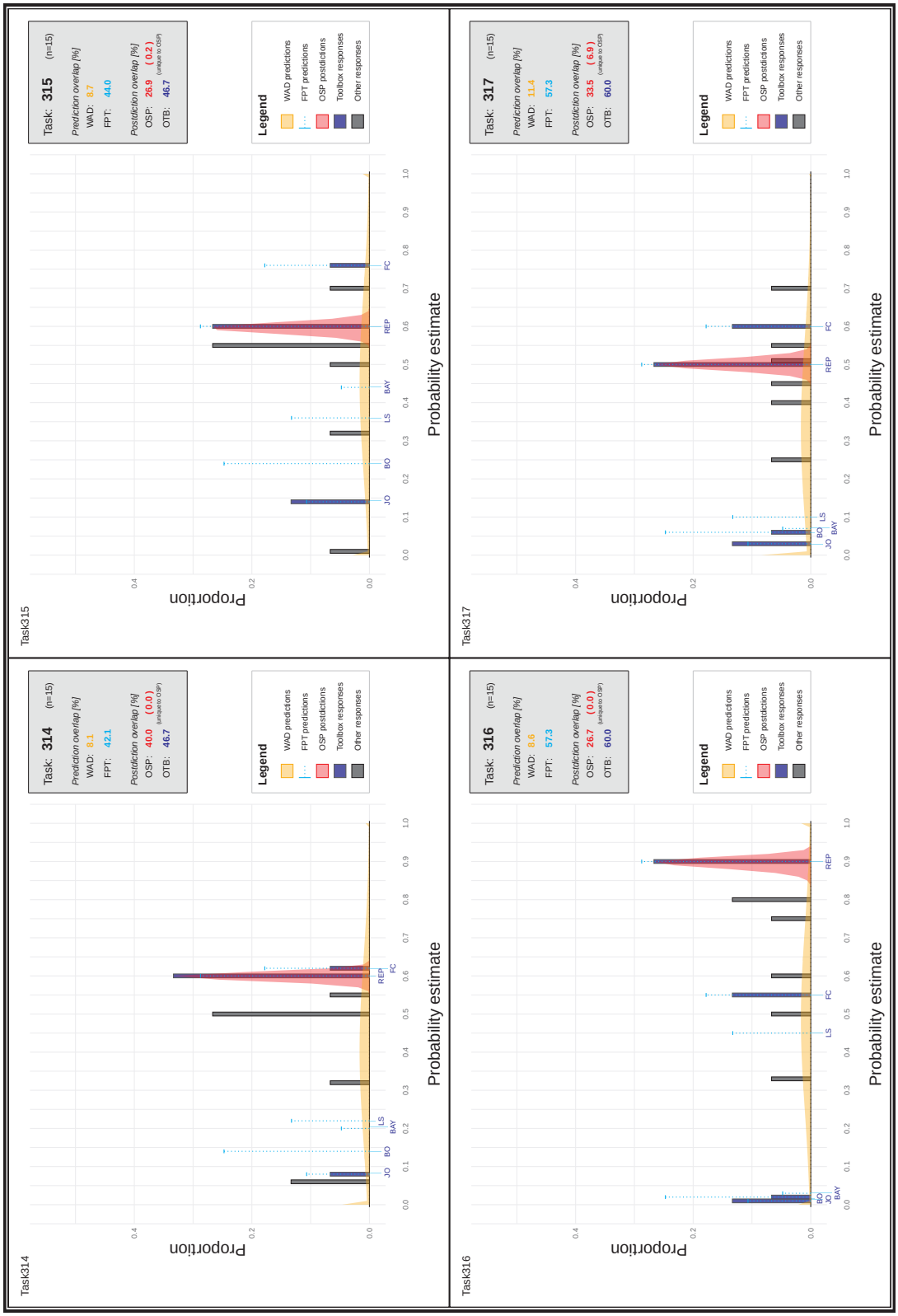

Figure S31. Comparison of toolbox and single-process predictions across tasks (part 9/27)

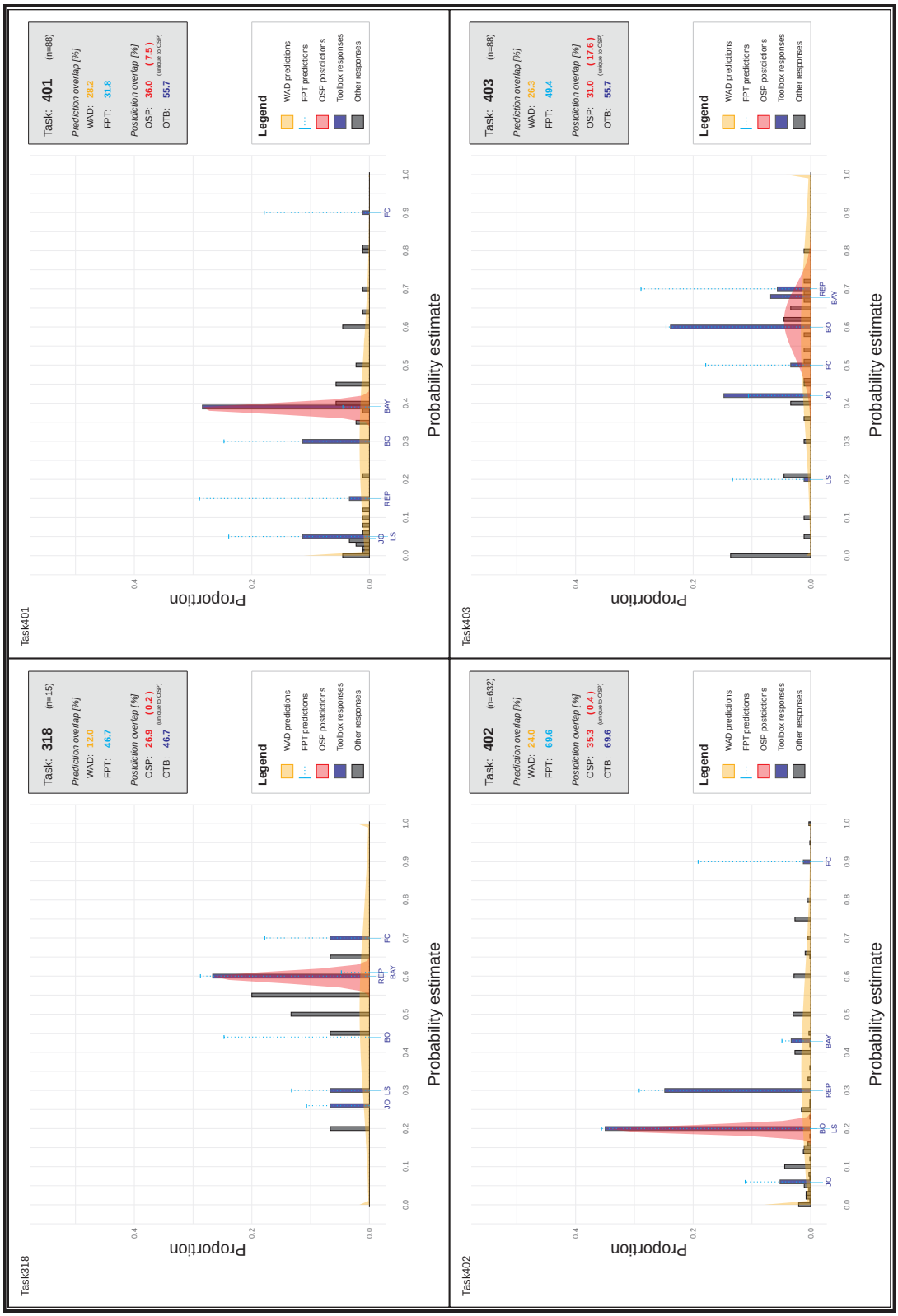

Figure S32. Comparison of toolbox and single-process predictions across tasks (part 10/27)

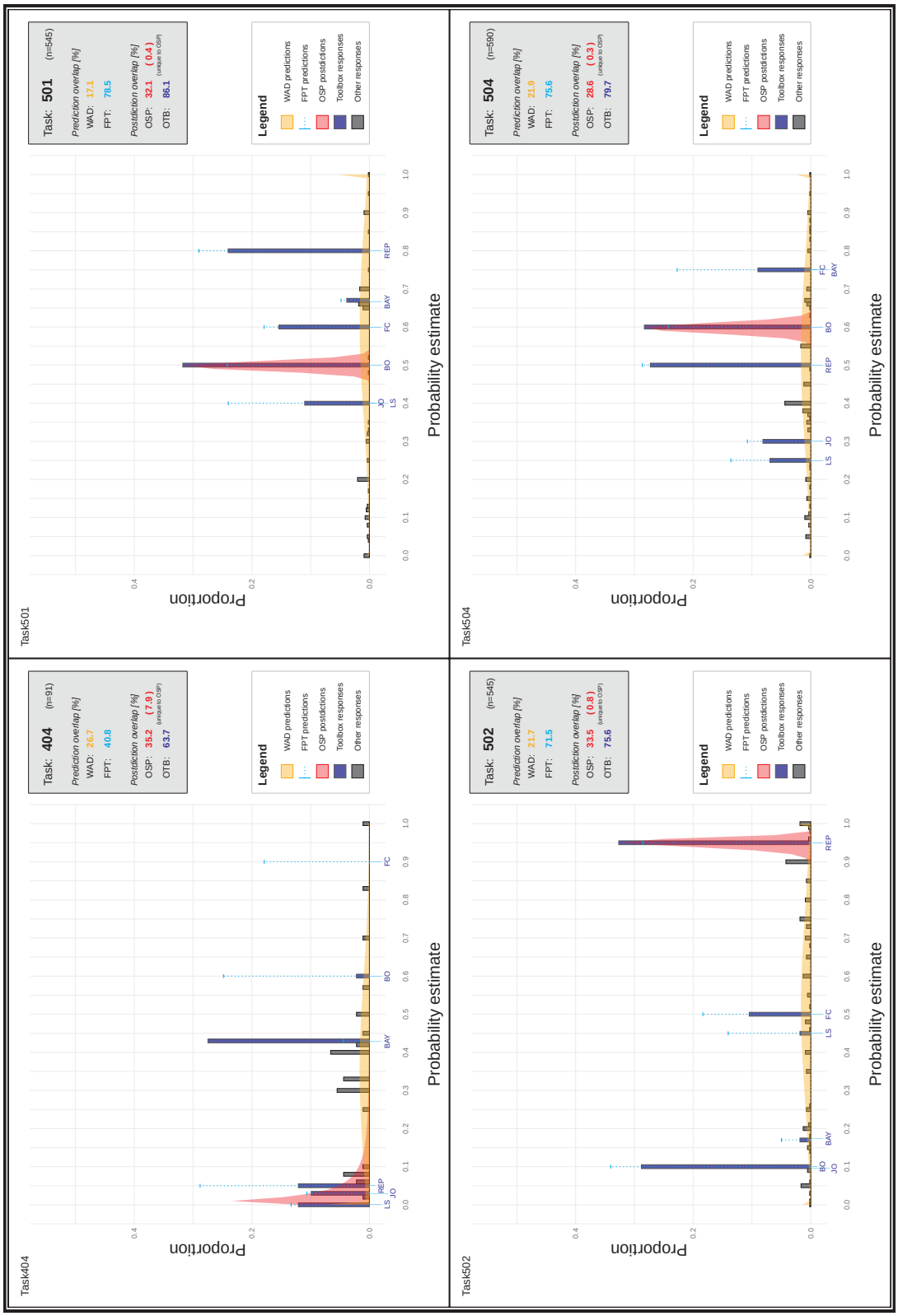

Figure S3. Comparison of toolbox and single-process predictions across tasks (part 11/27)

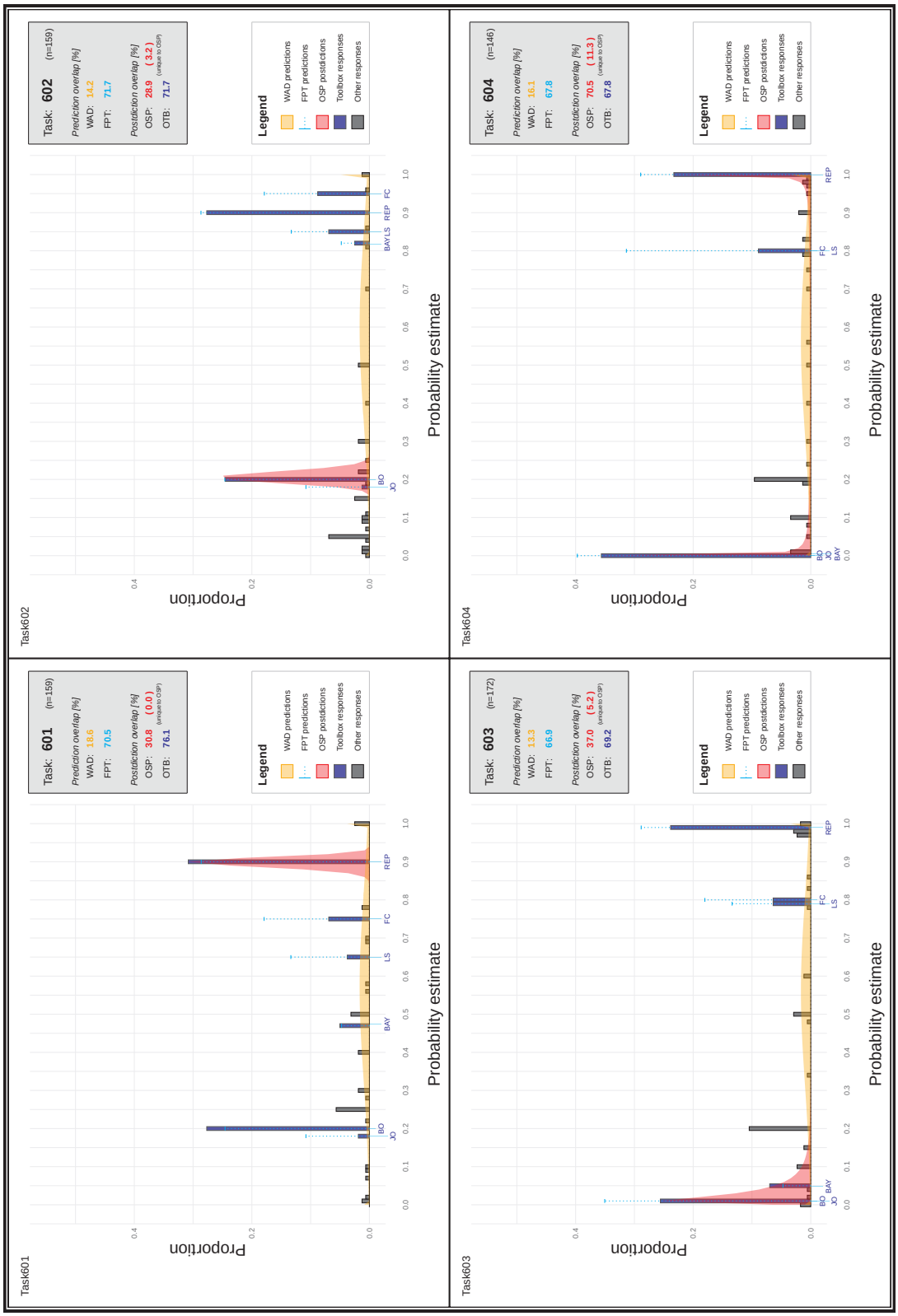

Figure S34. Comparison of toolbox and single-process predictions across tasks (part 12/27)

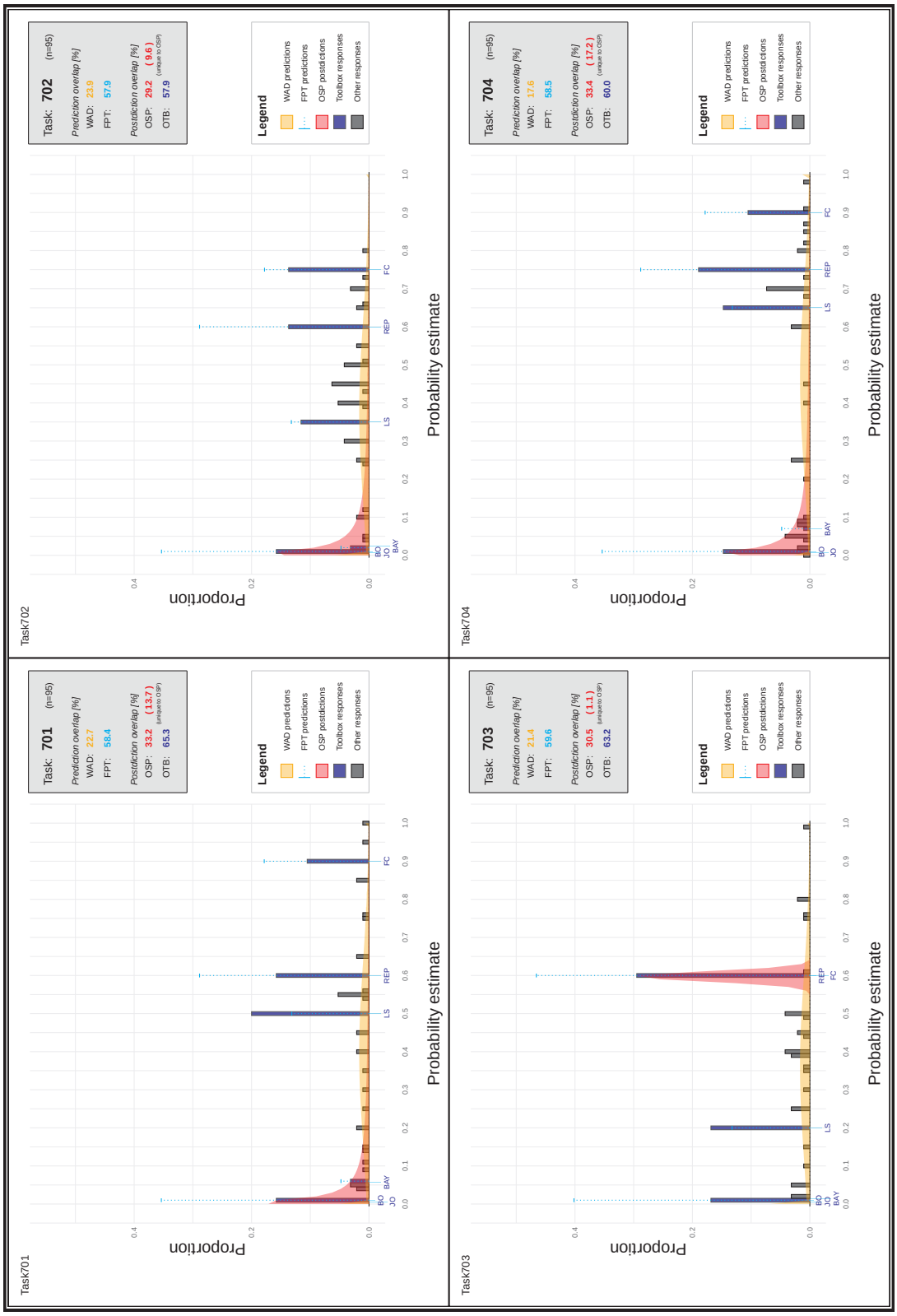

Figure S35. Comparison of toolbox and single-process predictions across tasks (part 13/27)

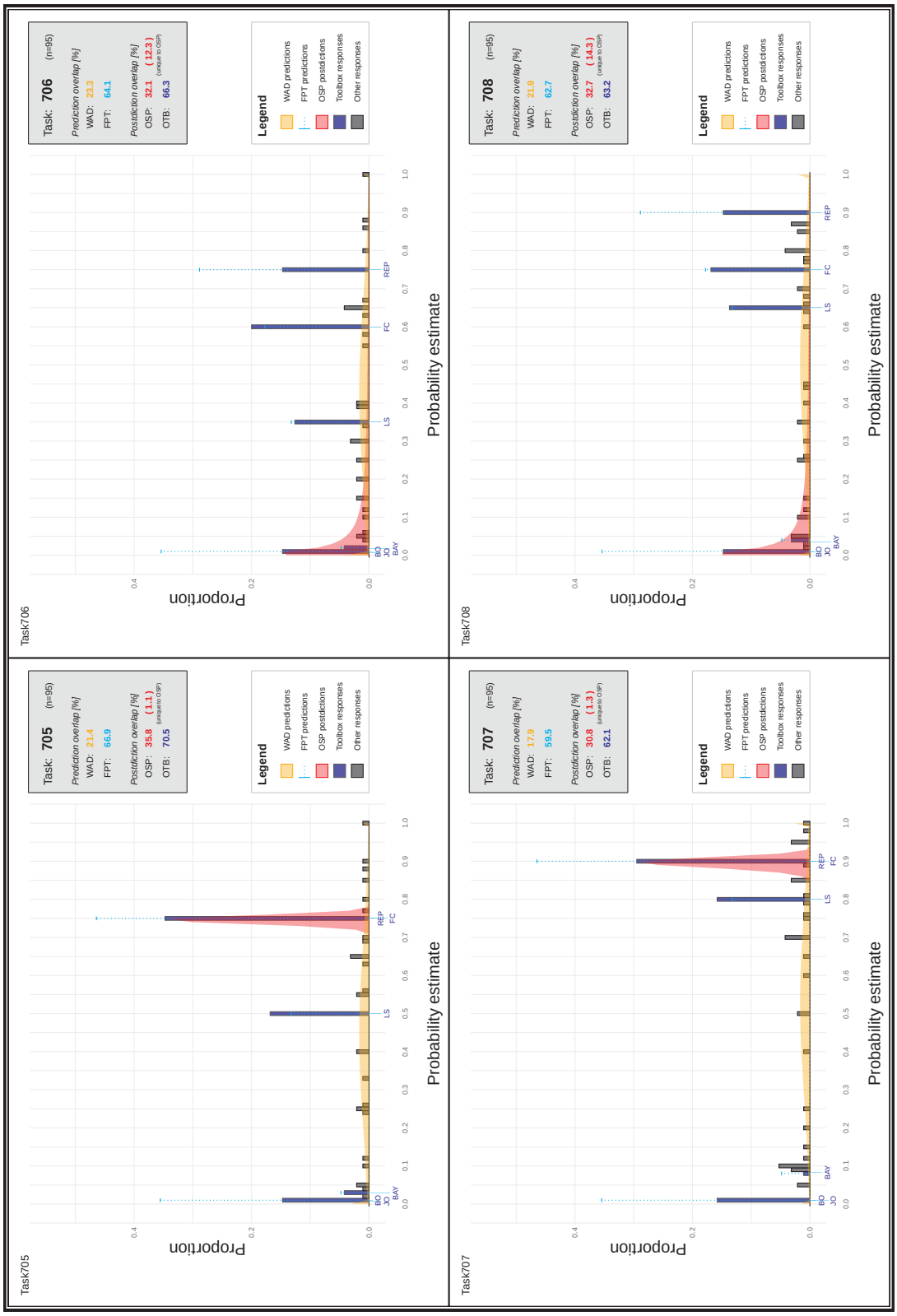

Figure S36. Comparison of toolbox and single-process predictions across tasks (part 14/27)

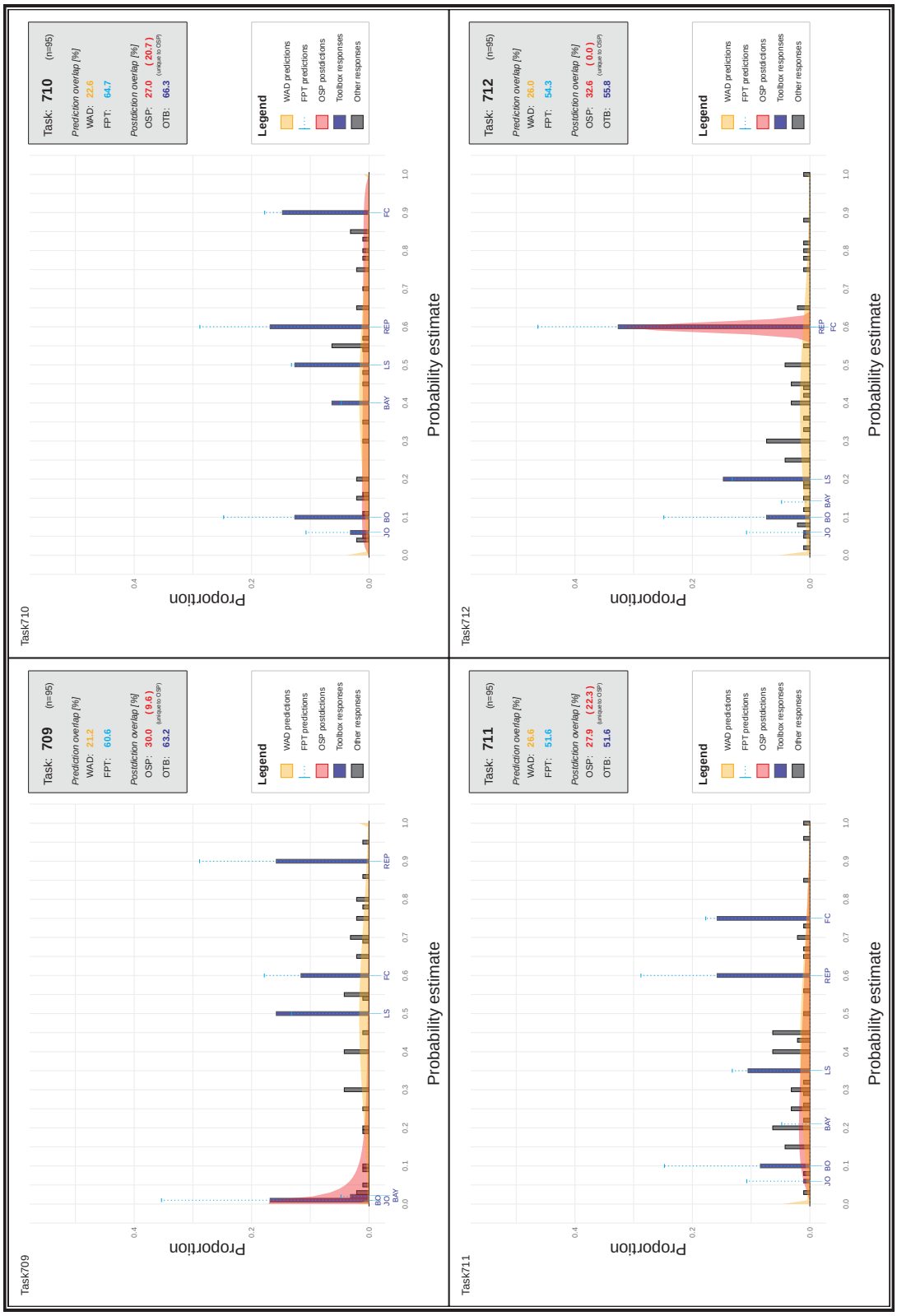

Figure S37. Comparison of toolbox and single-process predictions across tasks (part 15/27)

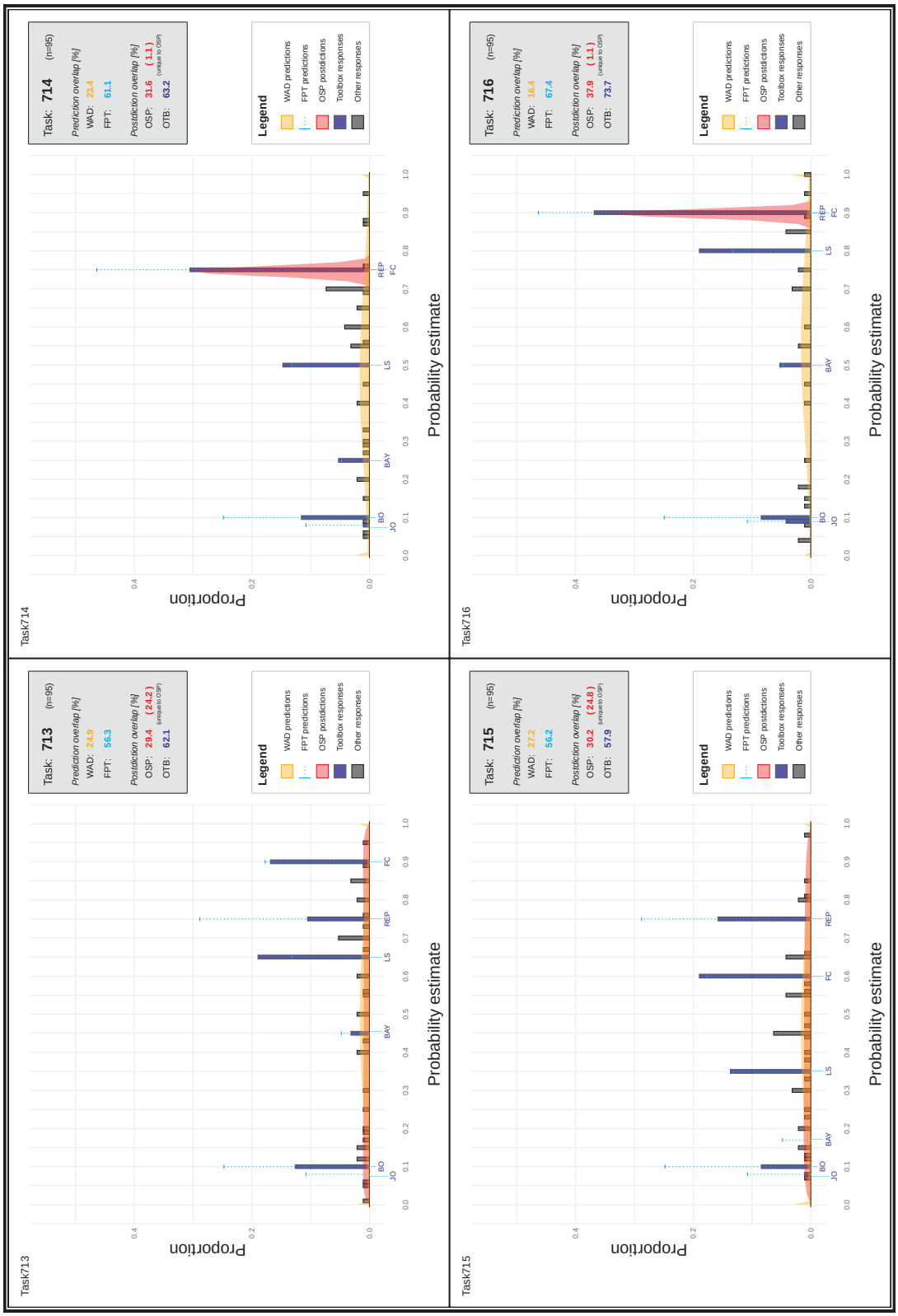

Figure S38. Comparison of toolbox and single-process predictions across tasks (part 16/27)

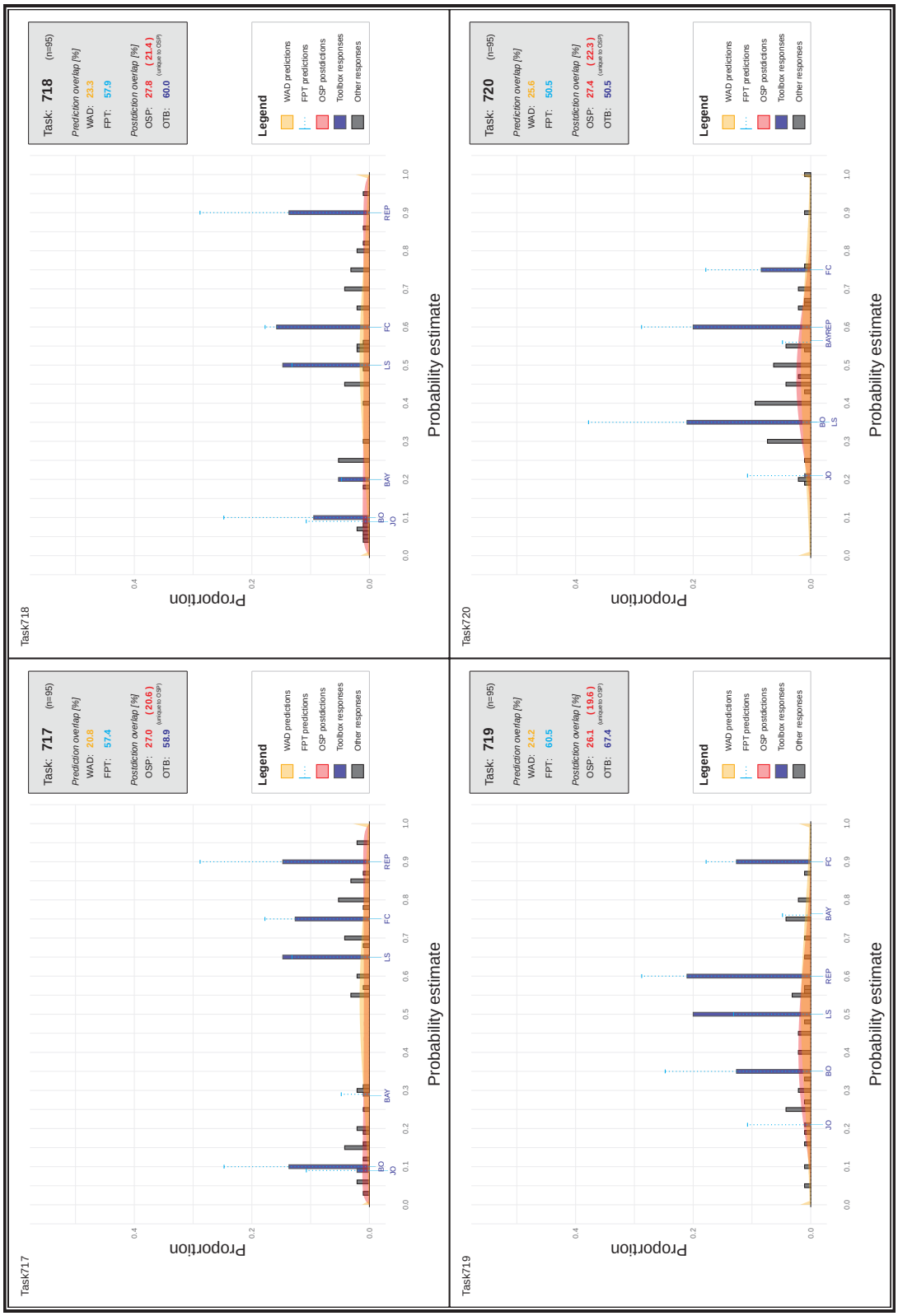

Figure S39. Comparison of toolbox and single-process predictions across tasks (part 17/27)

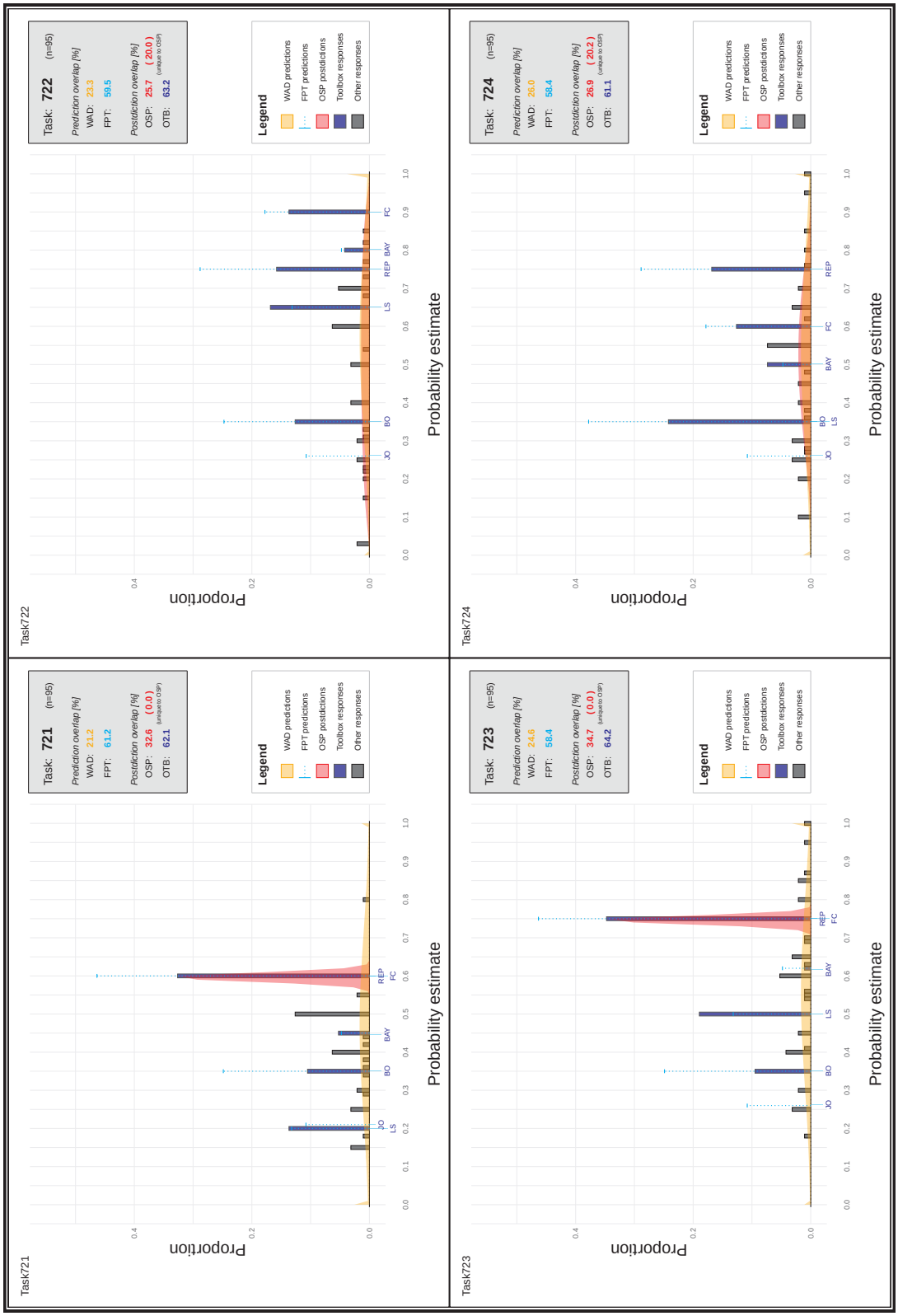

Figure S40. Comparison of toolbox and single-process predictions across tasks (part 18/27)

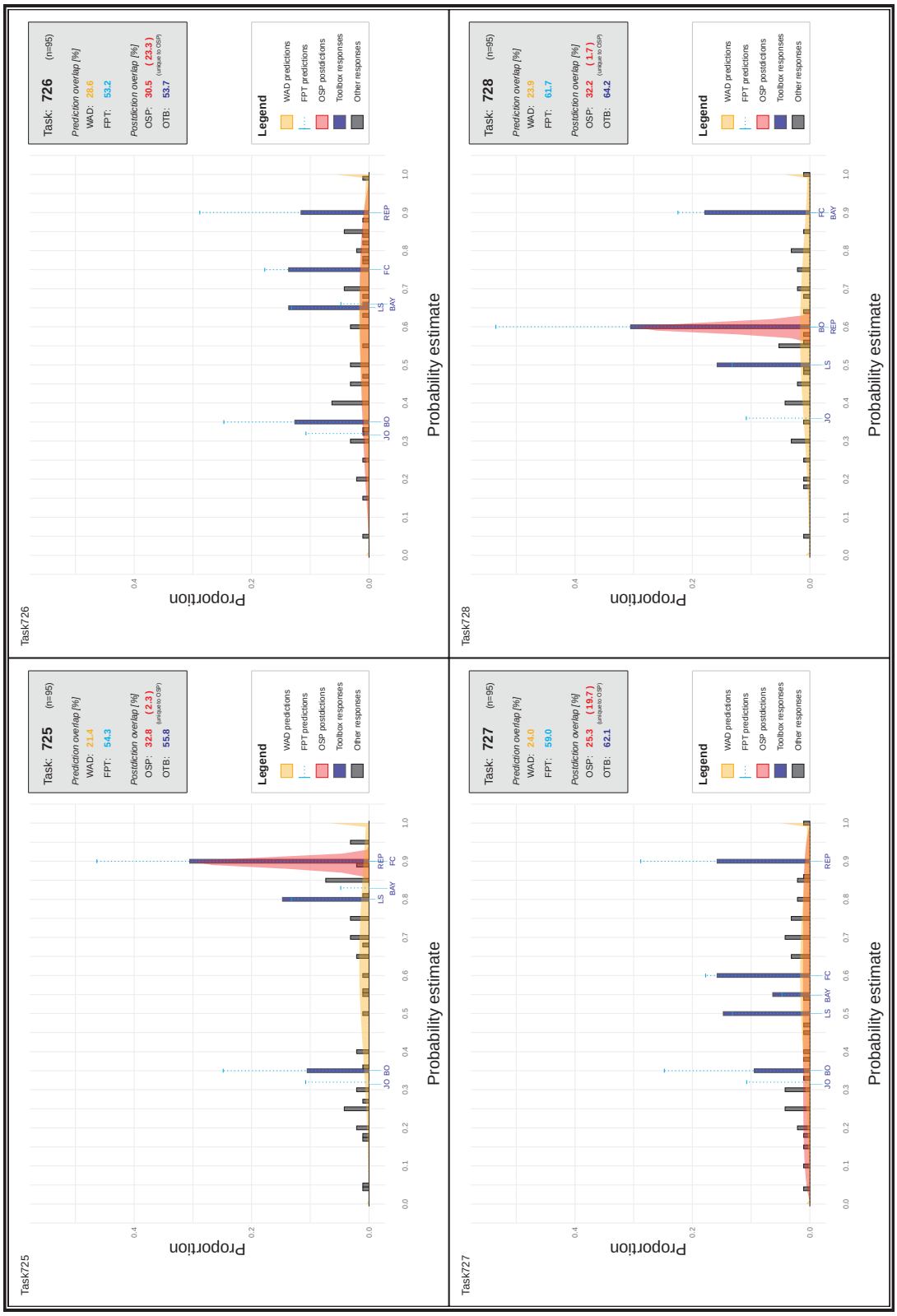

Figure S41. Comparison of toolbox and single-process predictions across tasks (part 19/27)

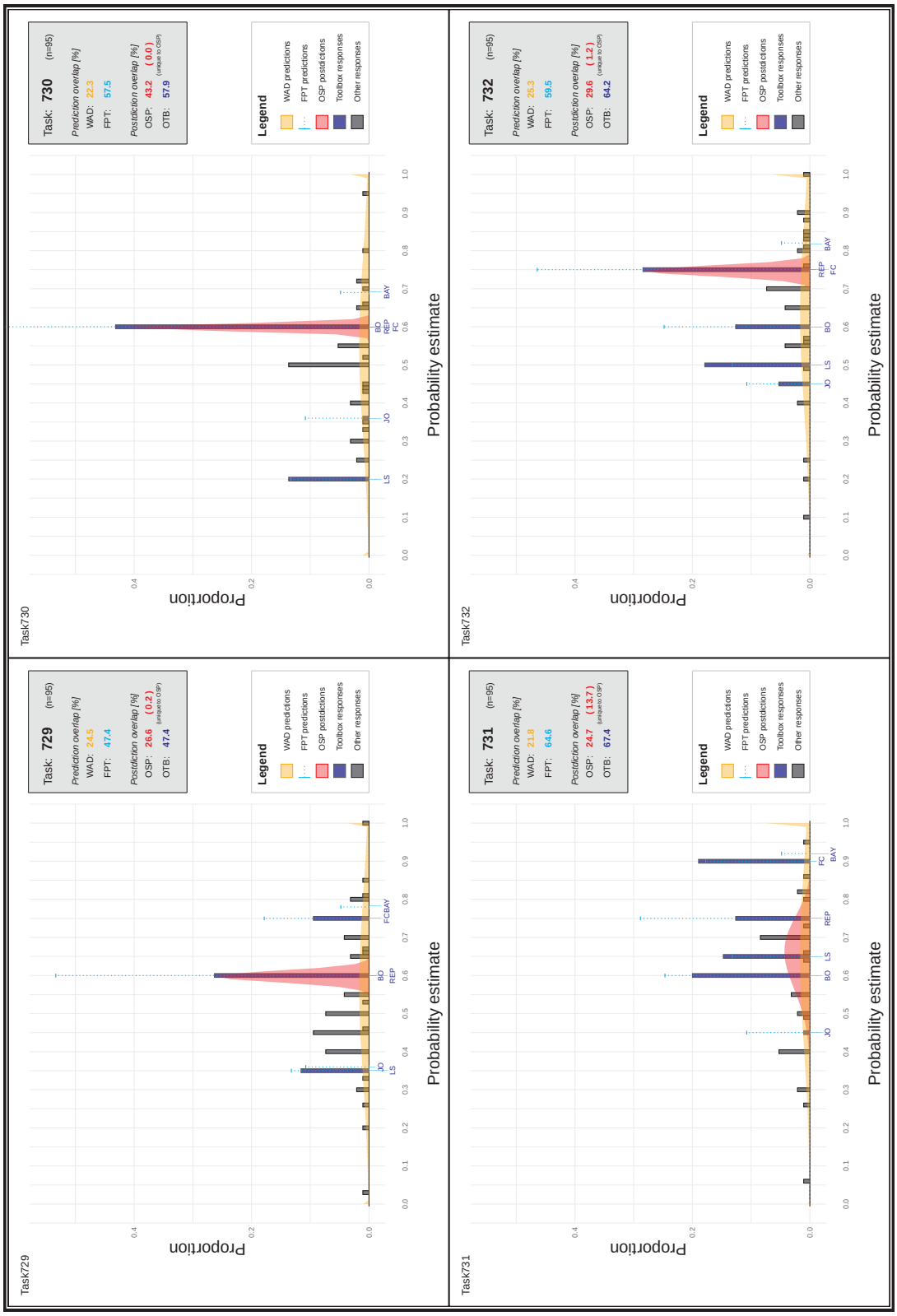

Figure S42. Comparison of toolbox and single-process predictions across tasks (part 20/27)

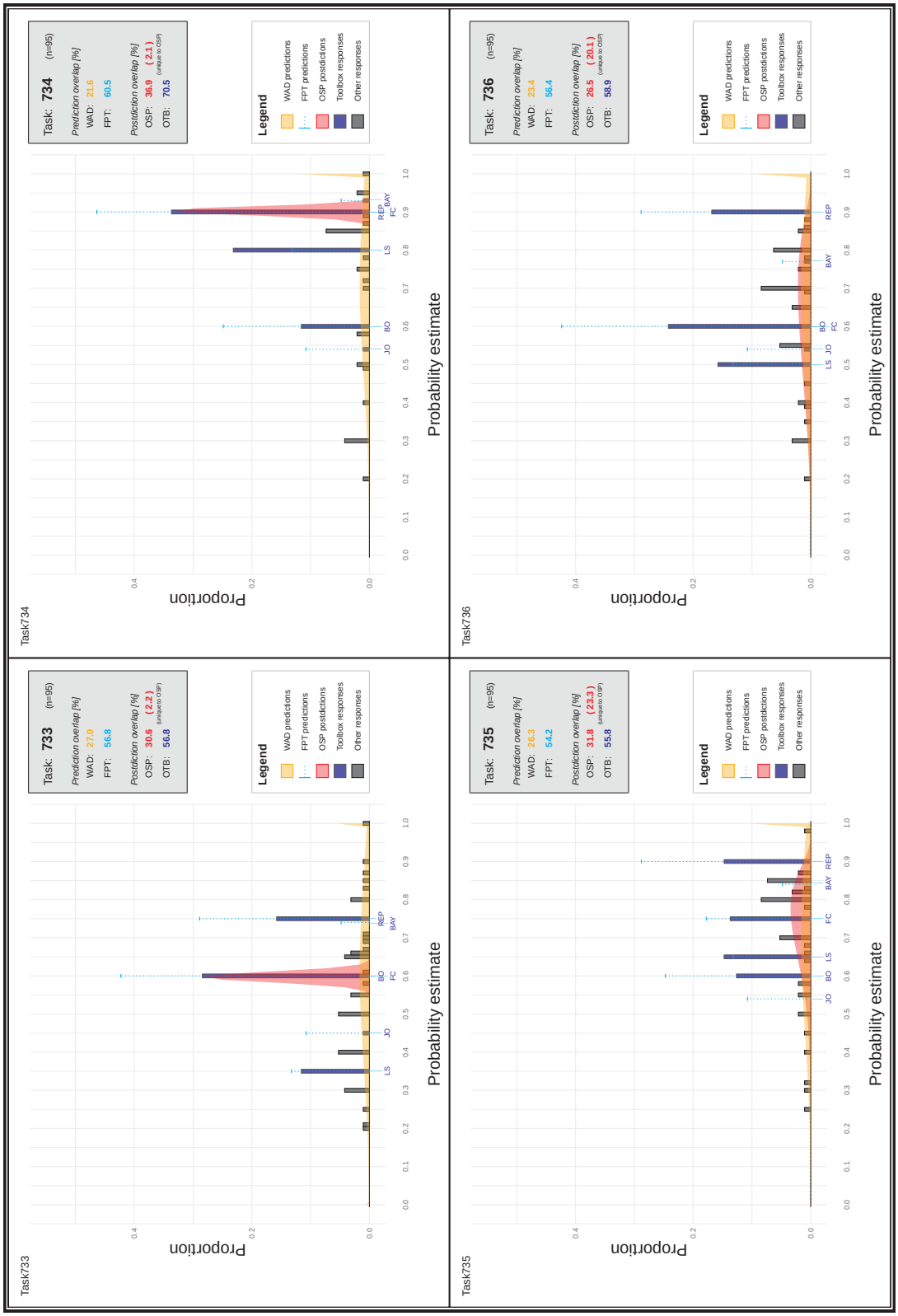

Figure S43. Comparison of toolbox and single-process predictions across tasks (part 21/27)

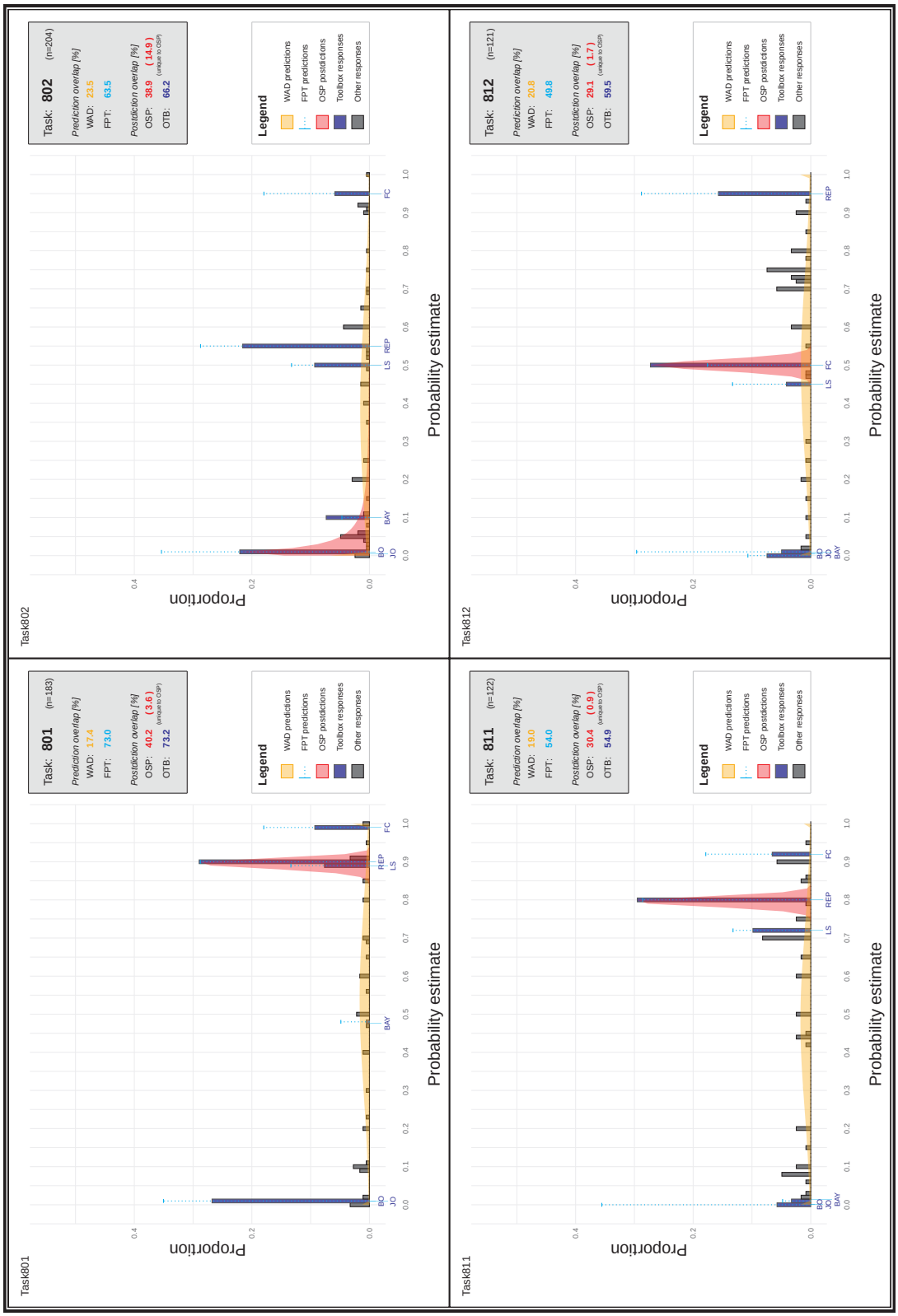

Figure S44. Comparison of toolbox and single-process predictions across tasks (part 22/27)

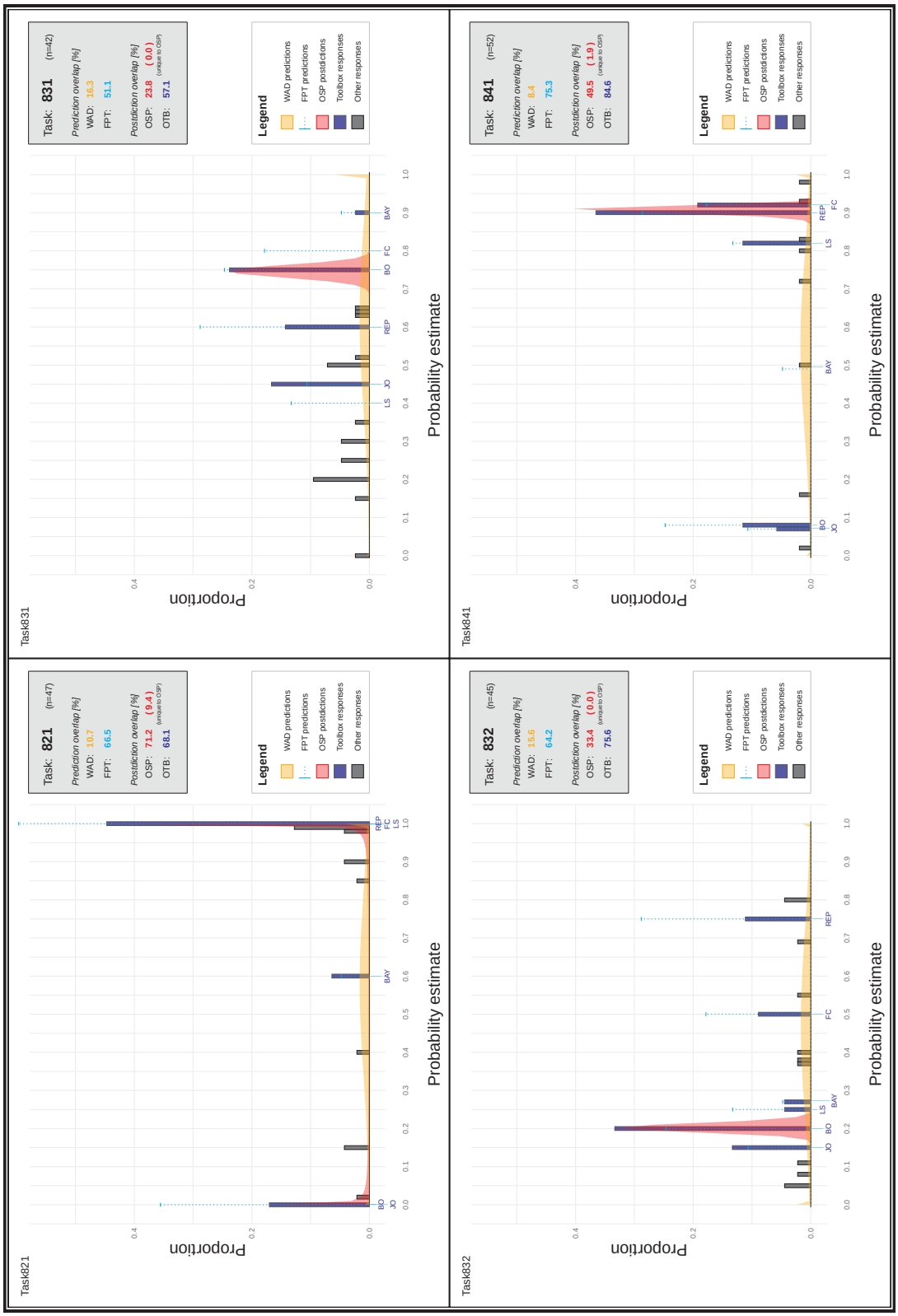

Figure S45. Comparison of toolbox and single-process predictions across tasks (part 23/27)

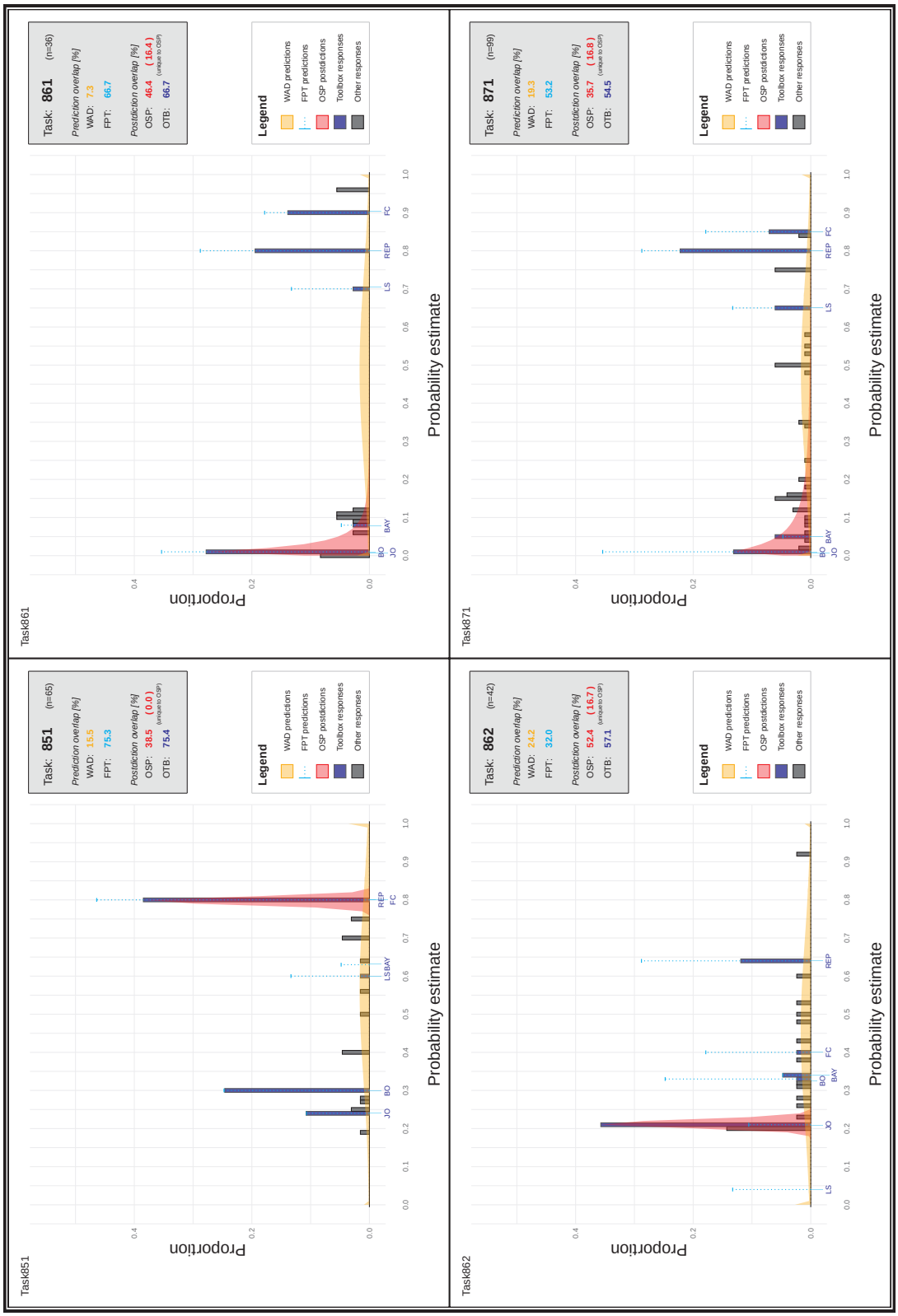

Figure S46. Comparison of toolbox and single-process predictions across tasks (part 24/27)

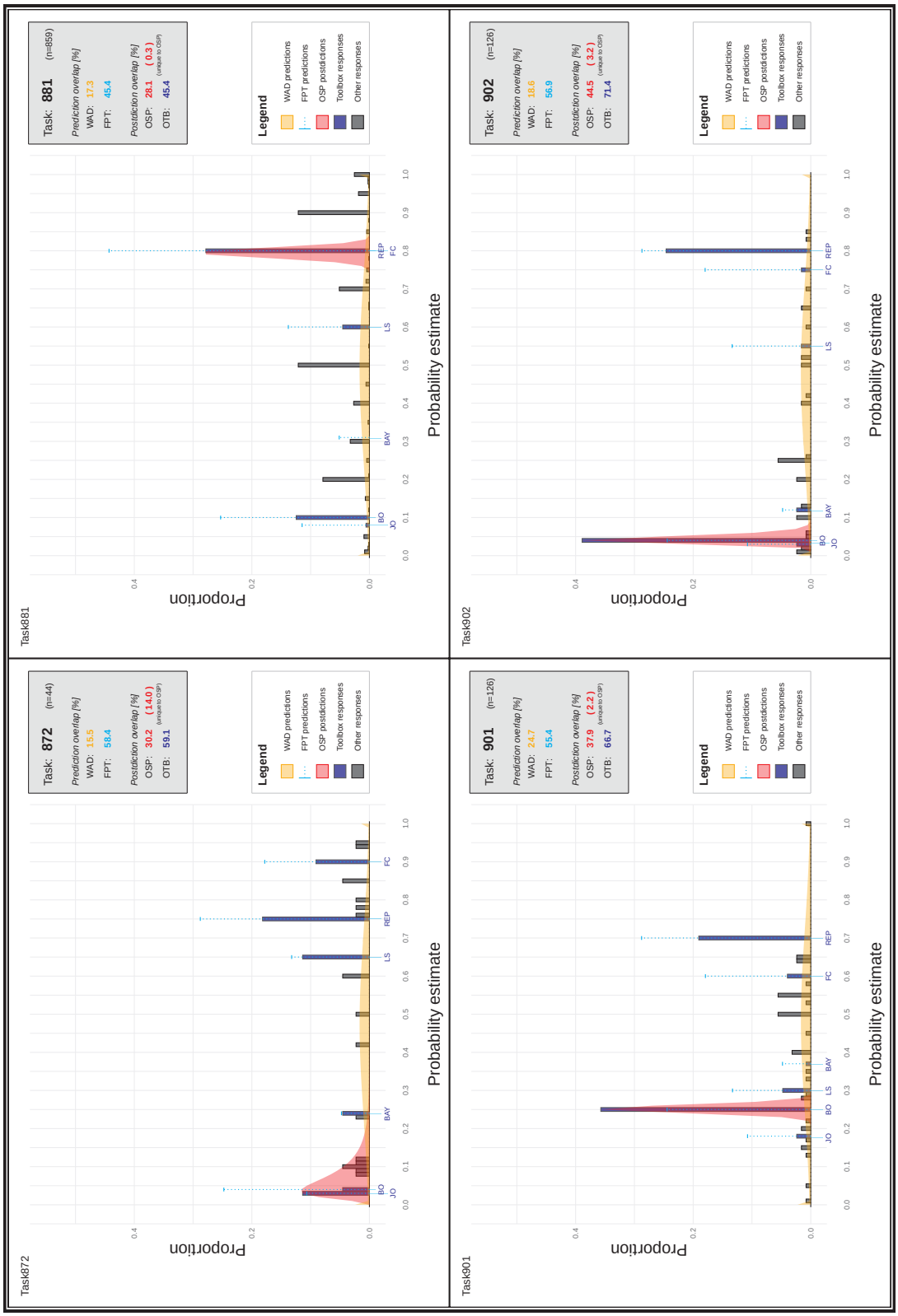

Figure S47. Comparison of toolbox and single-process predictions across tasks (part 25/27)

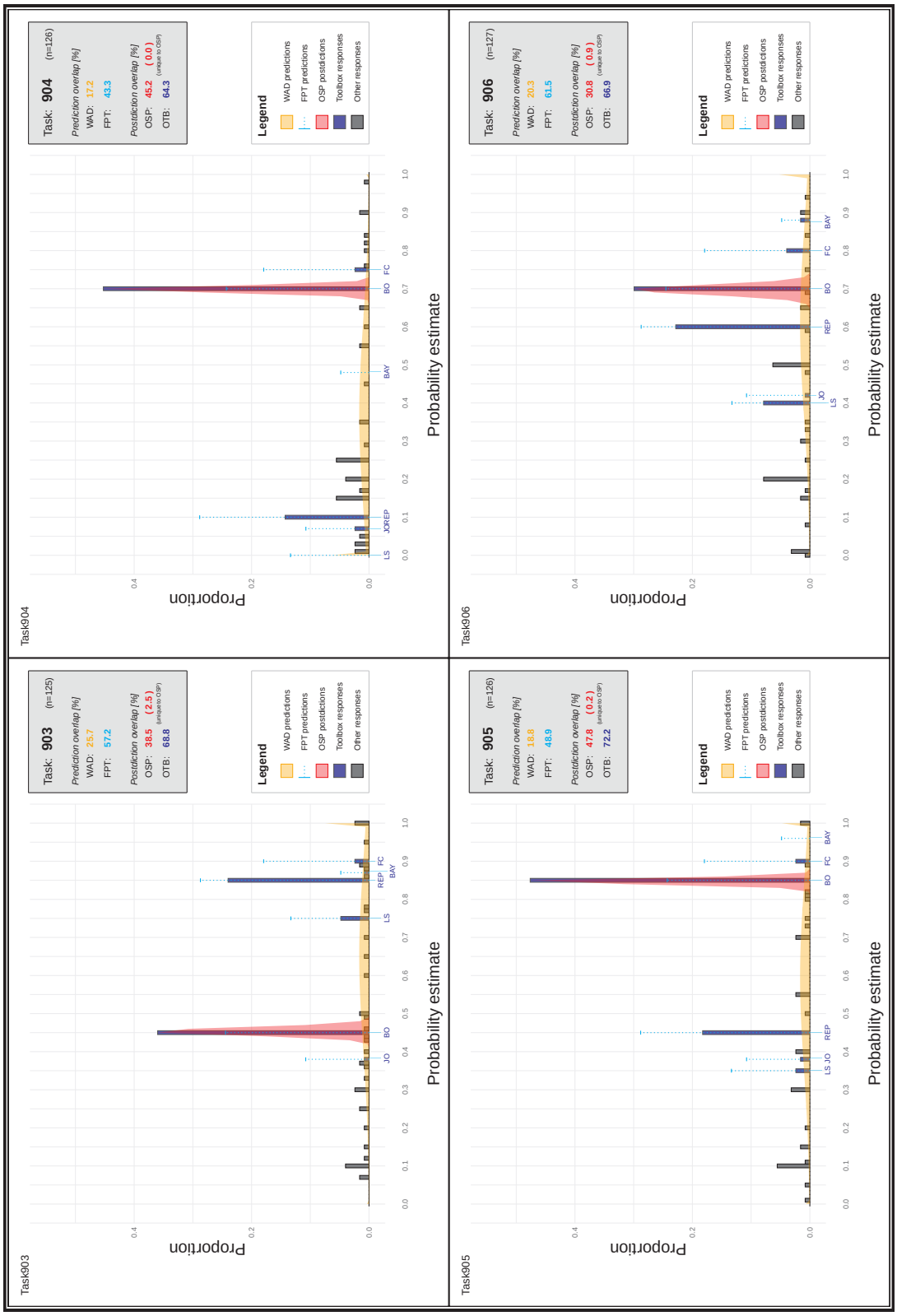

Figure S48. Comparison of toolbox and single-process predictions across tasks (part 26/27)

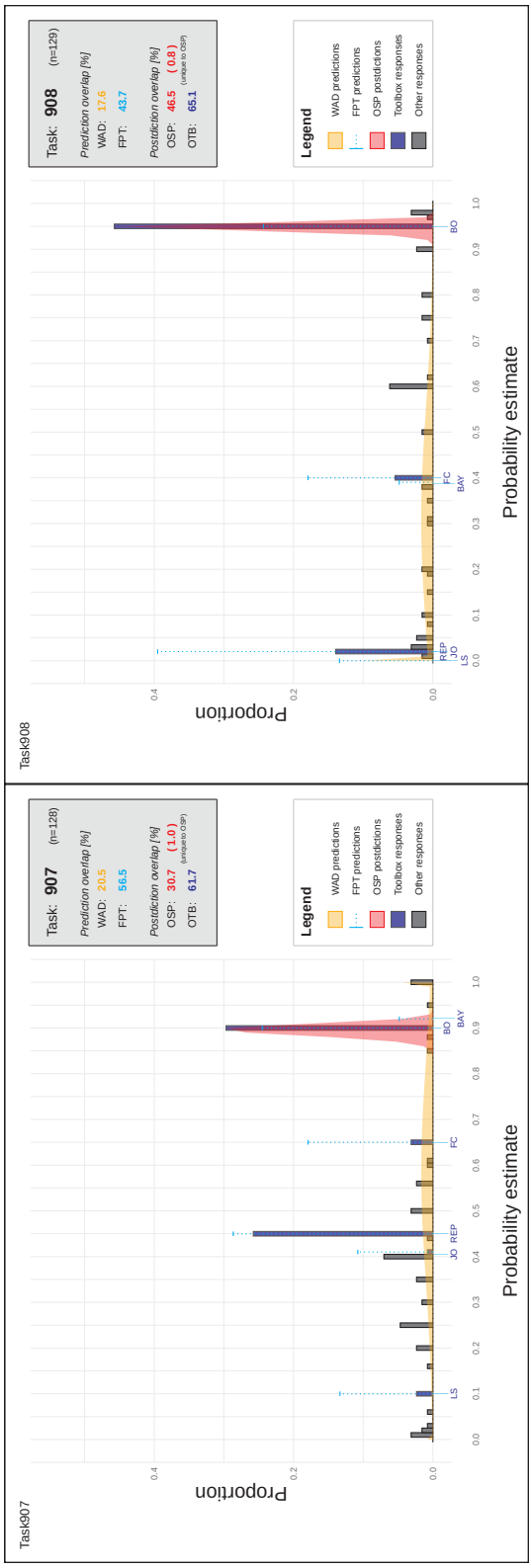

Figure S49. Comparison of toolbox and single-process predictions across tasks (part 27/27)

6.4.2 Comparison of conservatism and toolbox predictions. Each plot shows a comparison between estimates and overlap of conservatism models and toolbox models for one task (four plots are summarized in one figure). The plots present the distribution of relative response frequencies across intervals, as well as the estimates of the optimal toolbox model (OTB, blue bars) and estimates of the predictive five-plus toolbox model (FPT, dashed lines with whiskers). Overlaid are estimates of the predictive conservatism model based on restricted datasets with simple errors (CNr) and the optimal conservatism model with log-odds errors (OCo). The dashboards in the upper right summarize task information, the overlap performance of each model, and the overlap performance of the OCo model across intervals not estimated by the OTB. The shaded area shows the interval between base rate and the Bayesian posterior.

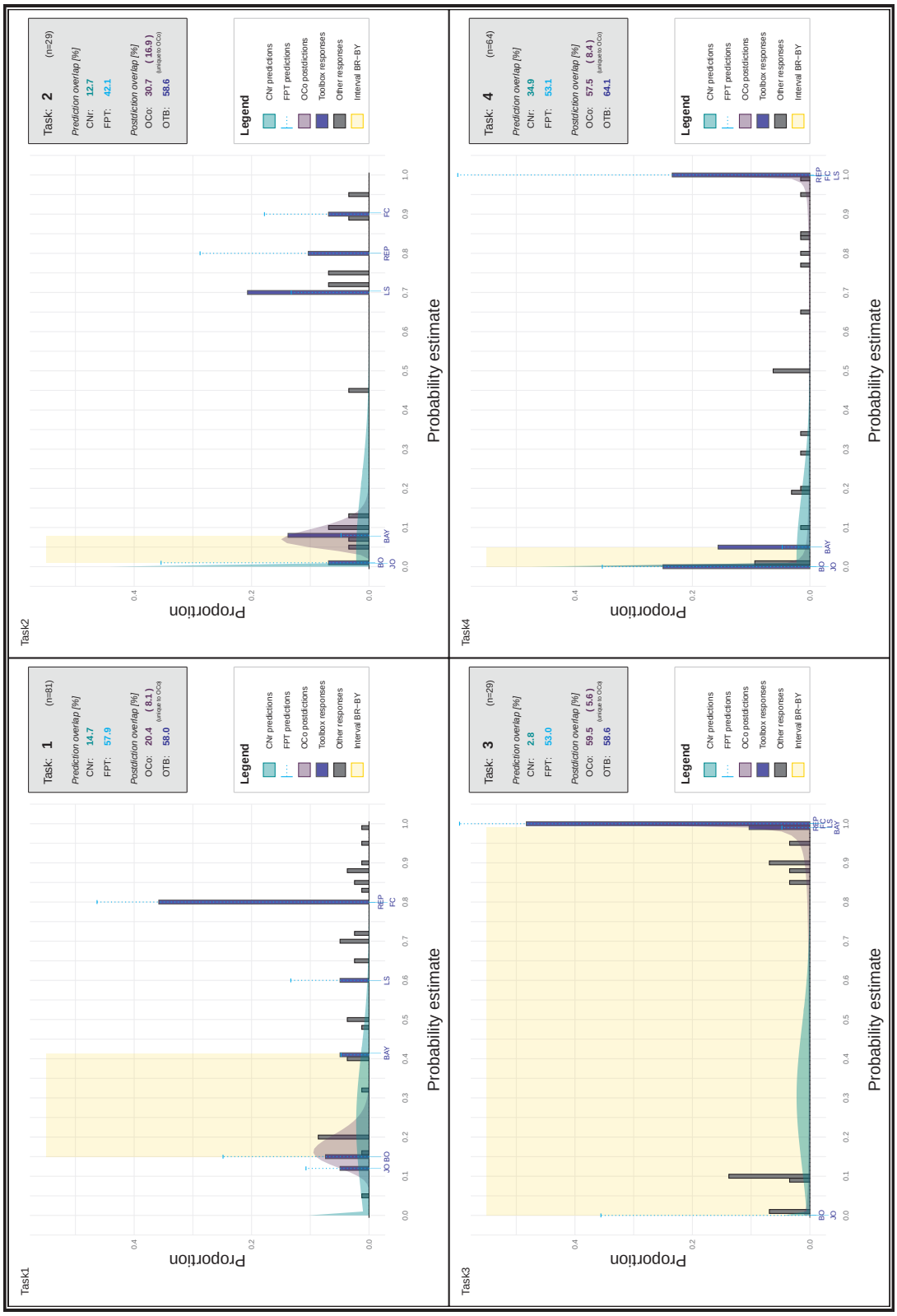

Figure S50. Comparison of toolbox and conservatism predictions across tasks (part 1/27)

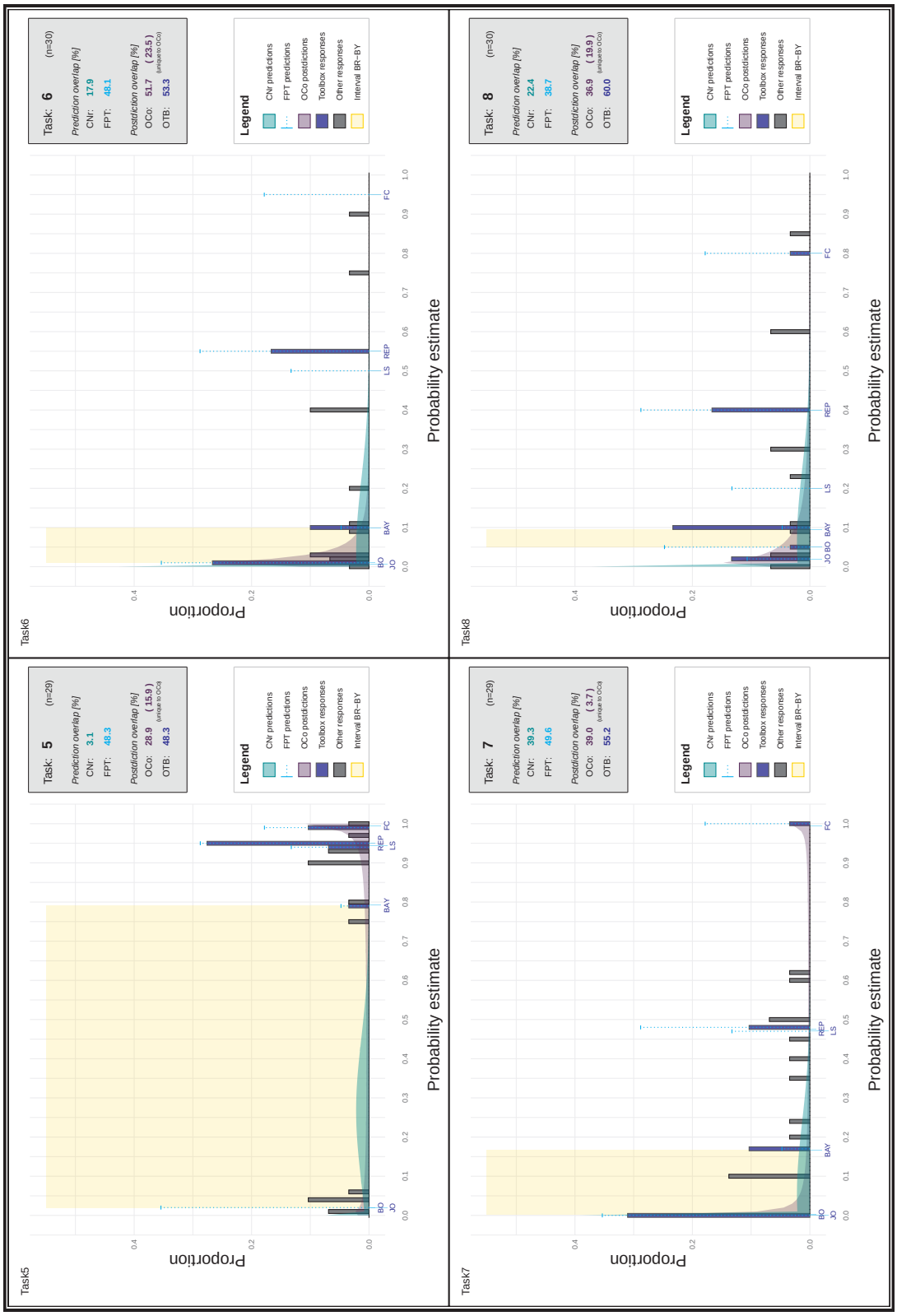

Figure S51. Comparison of toolbox and conservatism predictions across tasks (part 2/27)

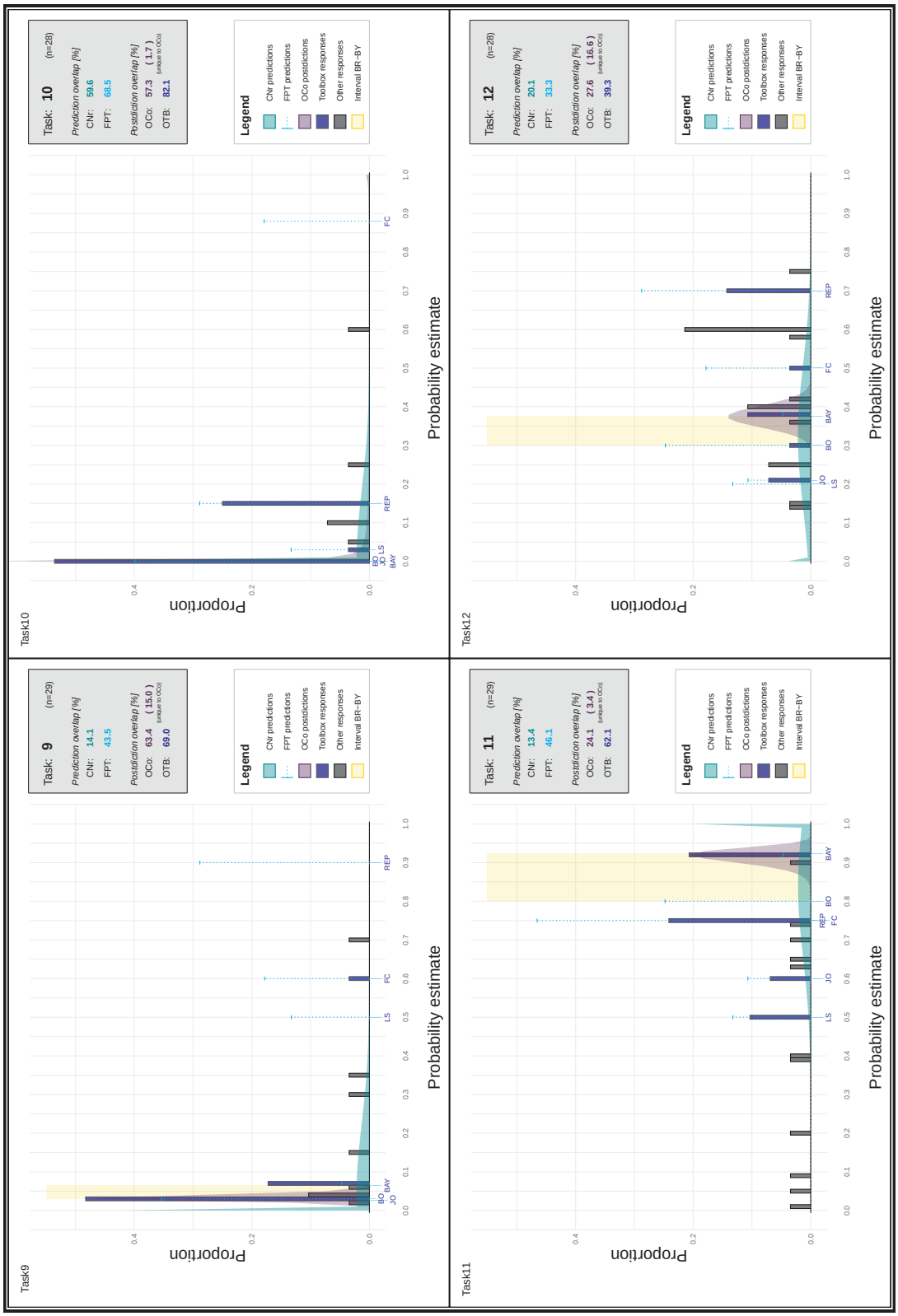

Figure S52. Comparison of toolbox and conservatism predictions across tasks (part 3/27)

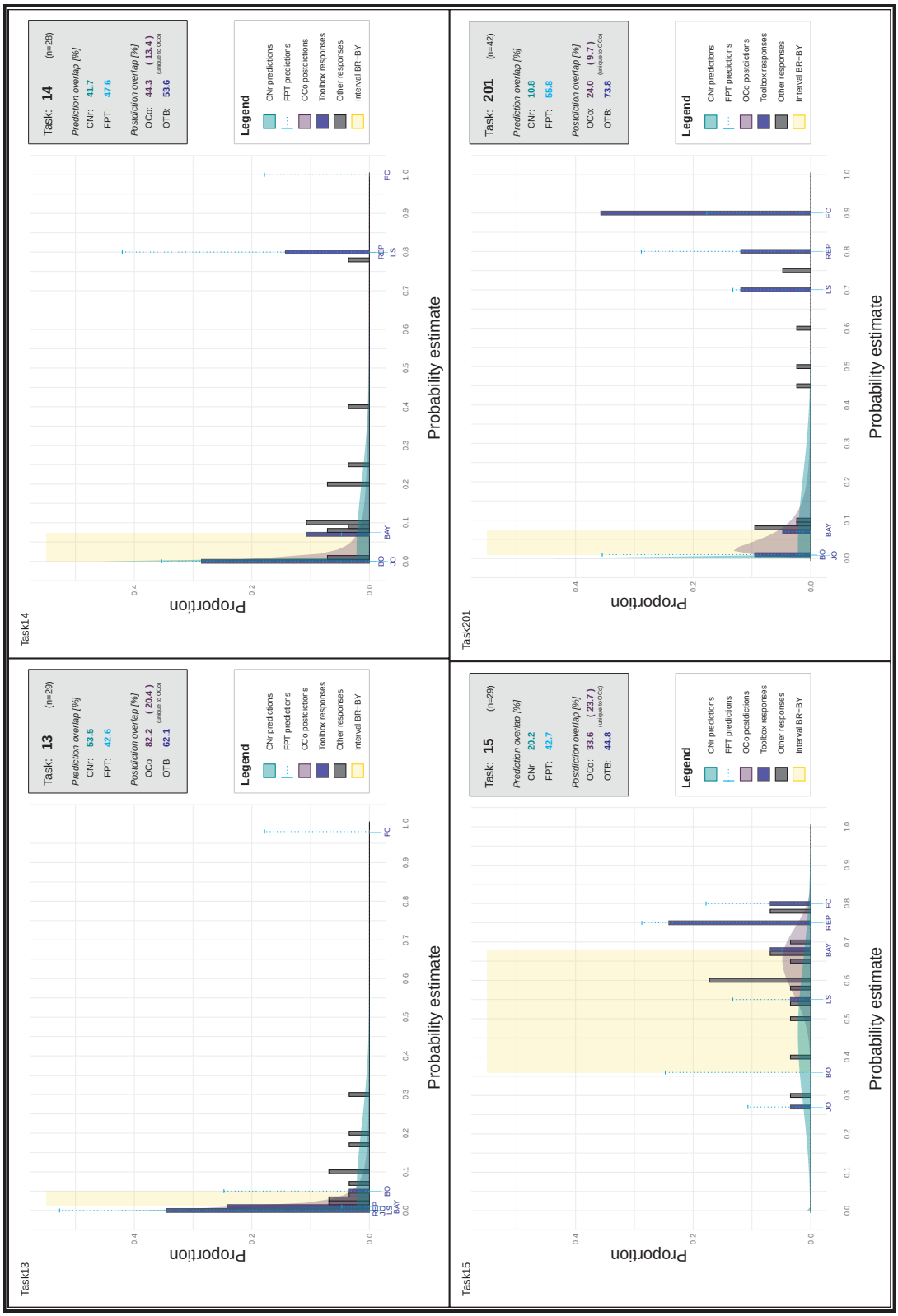

Figure S53. Comparison of toolbox and conservatism predictions across tasks (part 4/27)

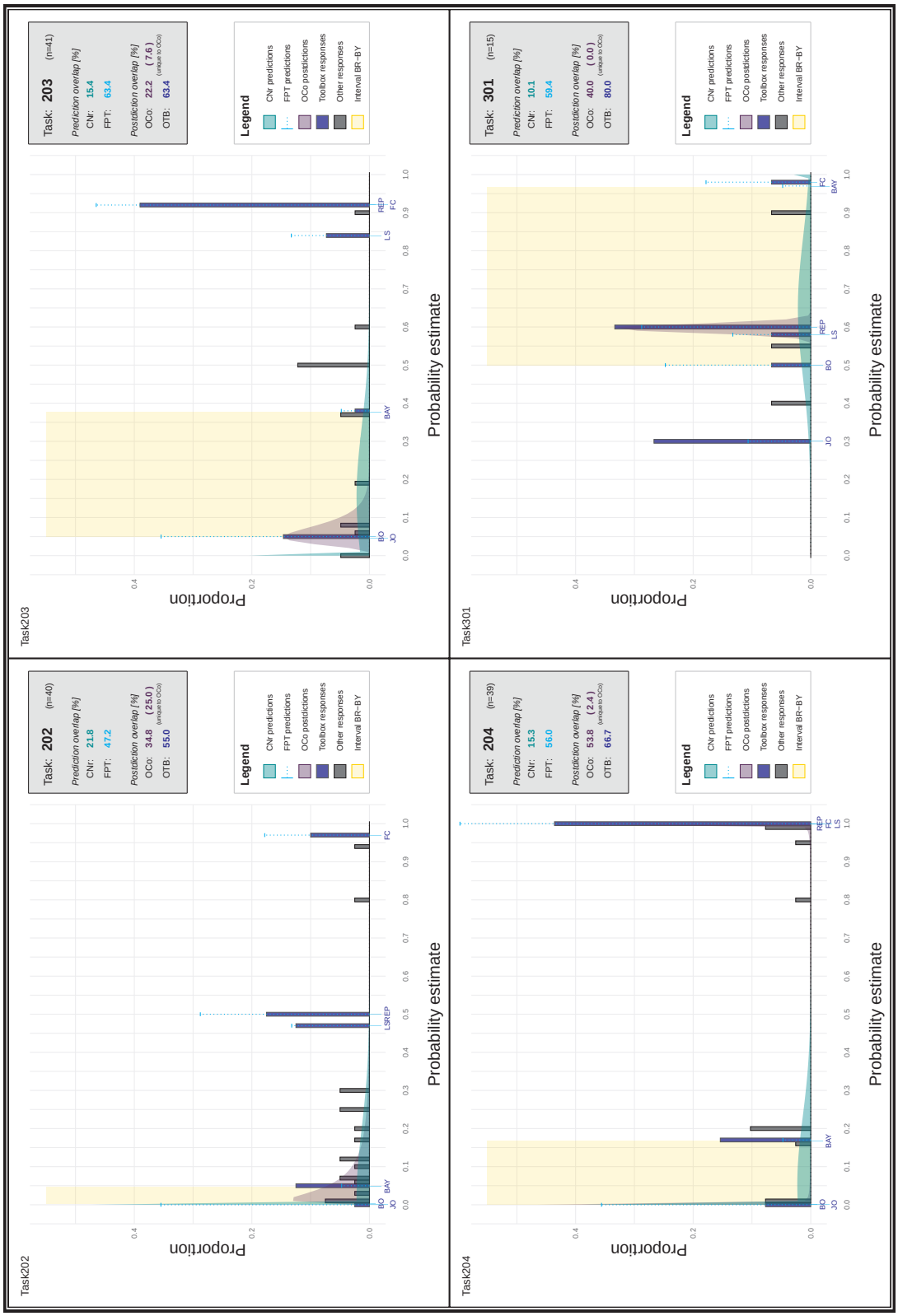

Figure S54. Comparison of toolbox and conservatism predictions across tasks (part 5/27)

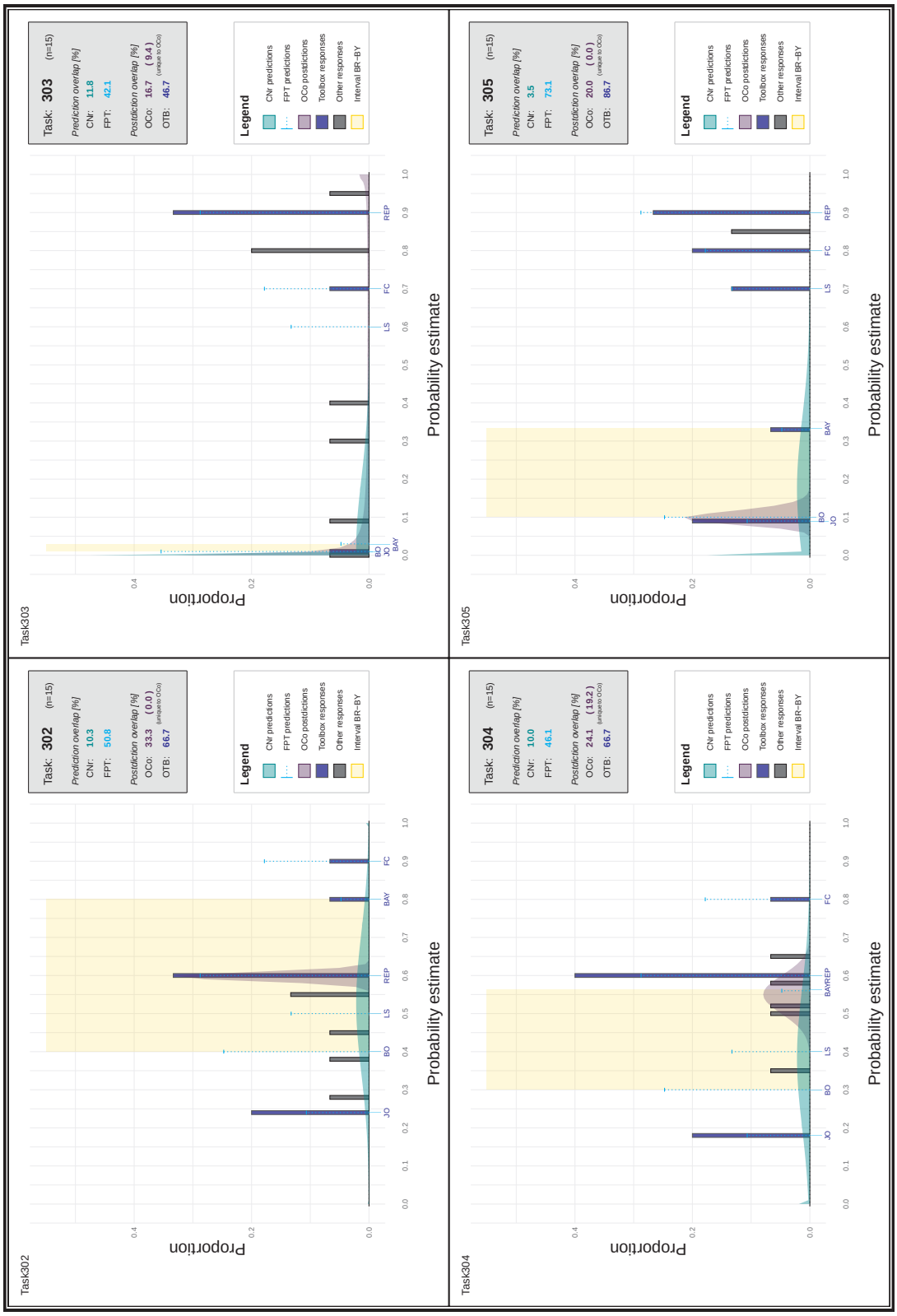

Figure S55. Comparison of toolbox and conservatism predictions across tasks (part 6/27)

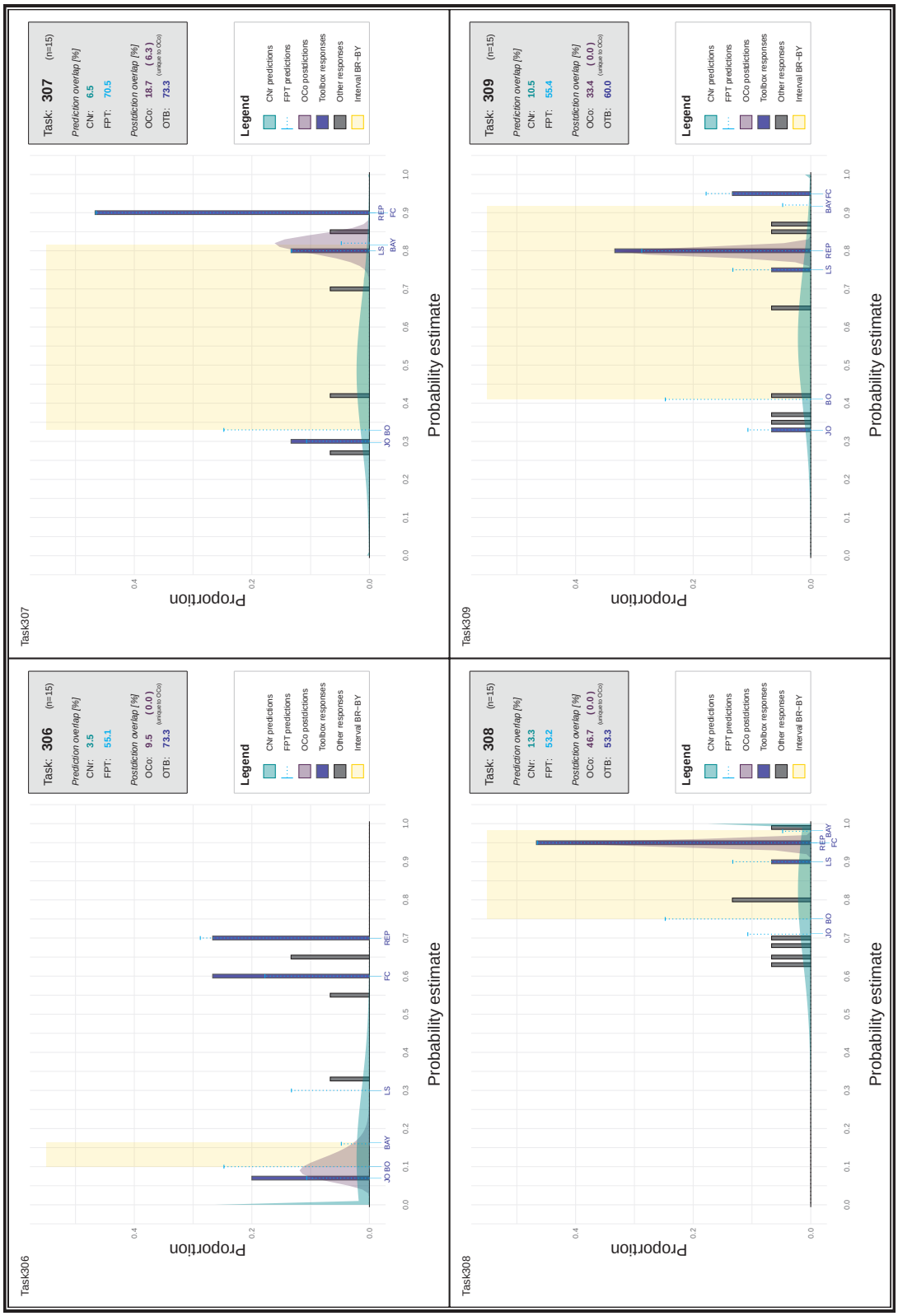

Figure S56. Comparison of toolbox and conservatism predictions across tasks (part 7/27)

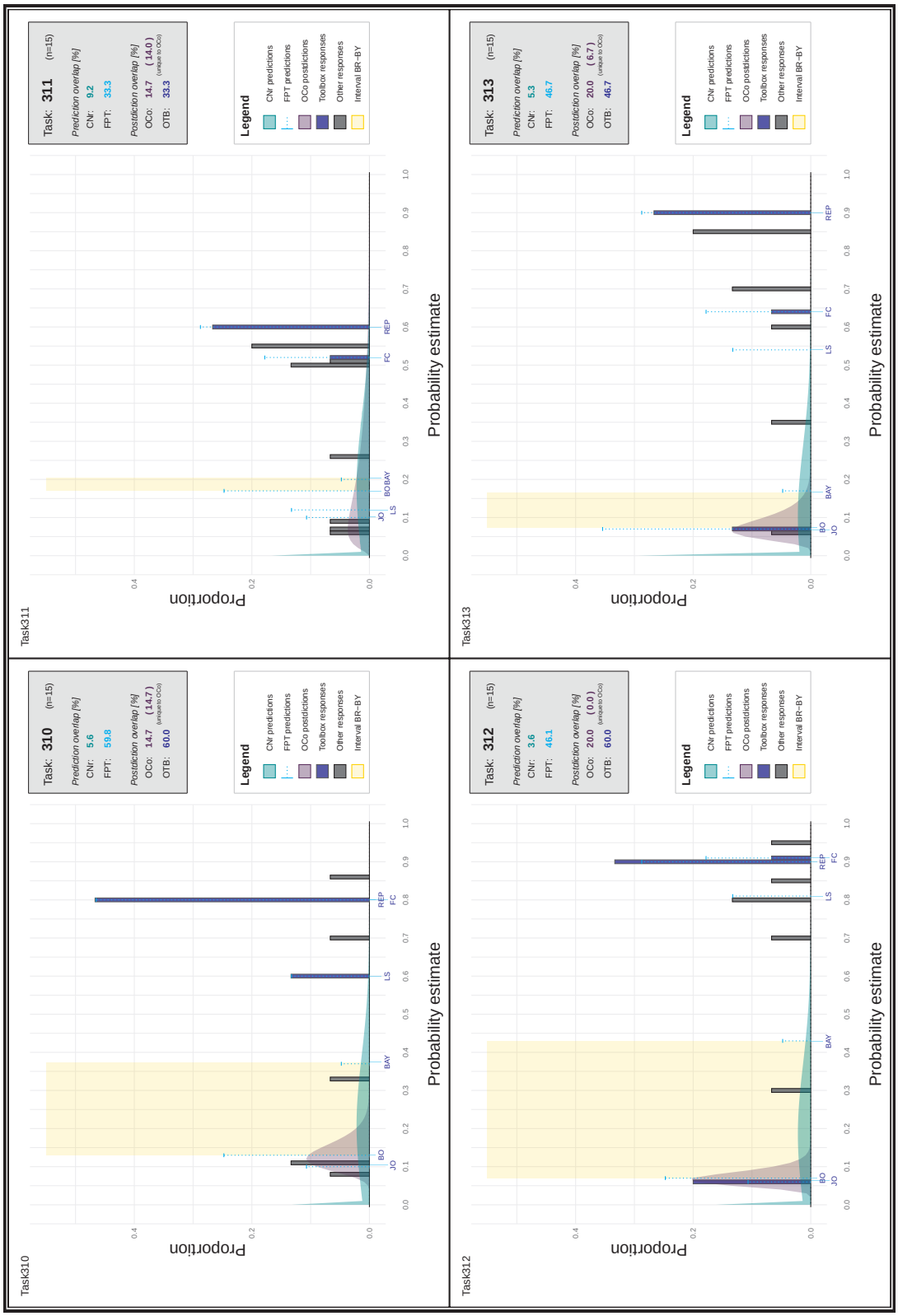

Figure S57. Comparison of toolbox and conservatism predictions across tasks (part 8/27)

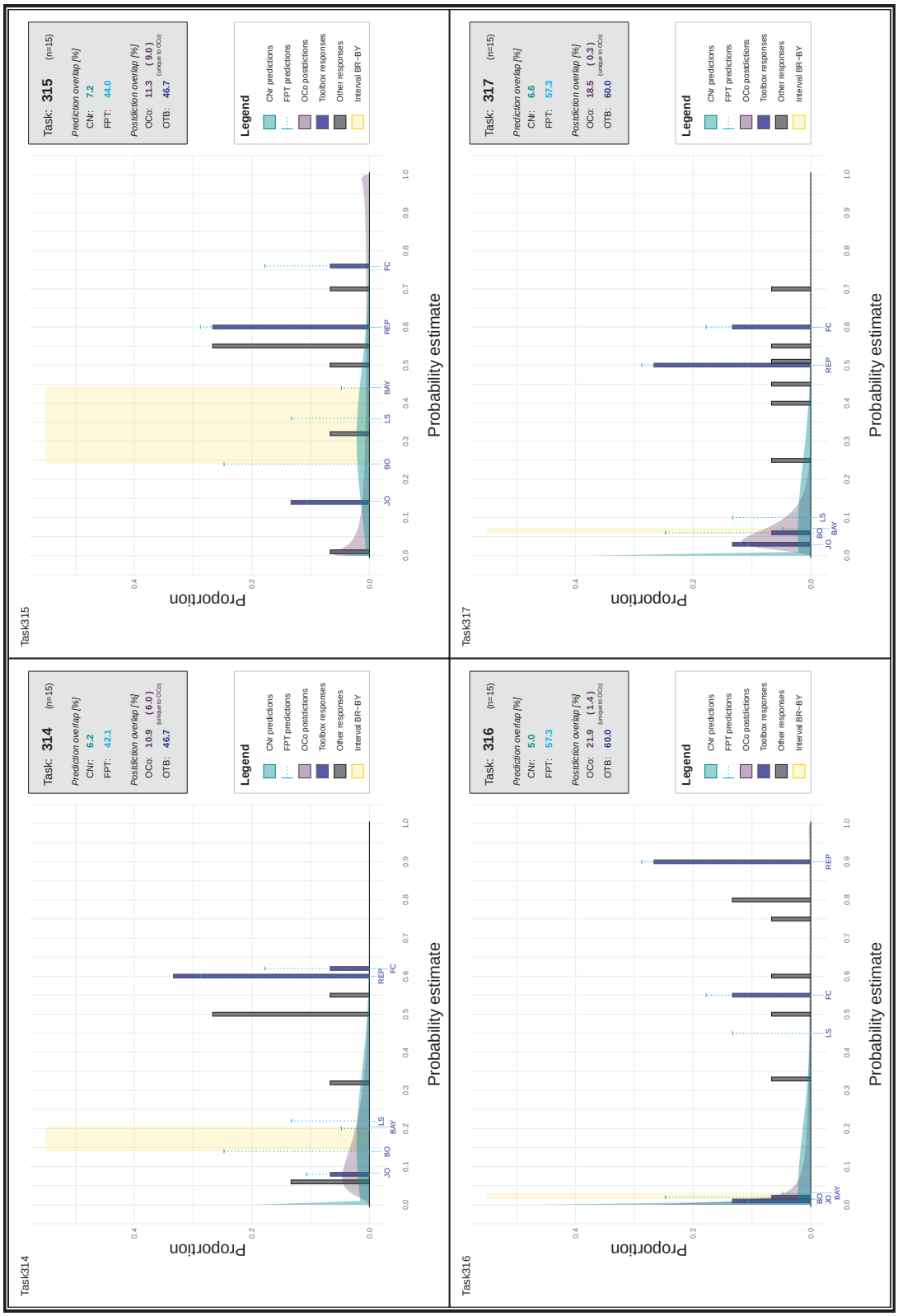

Figure S58. Comparison of toolbox and conservatism predictions across tasks (part 9/27)

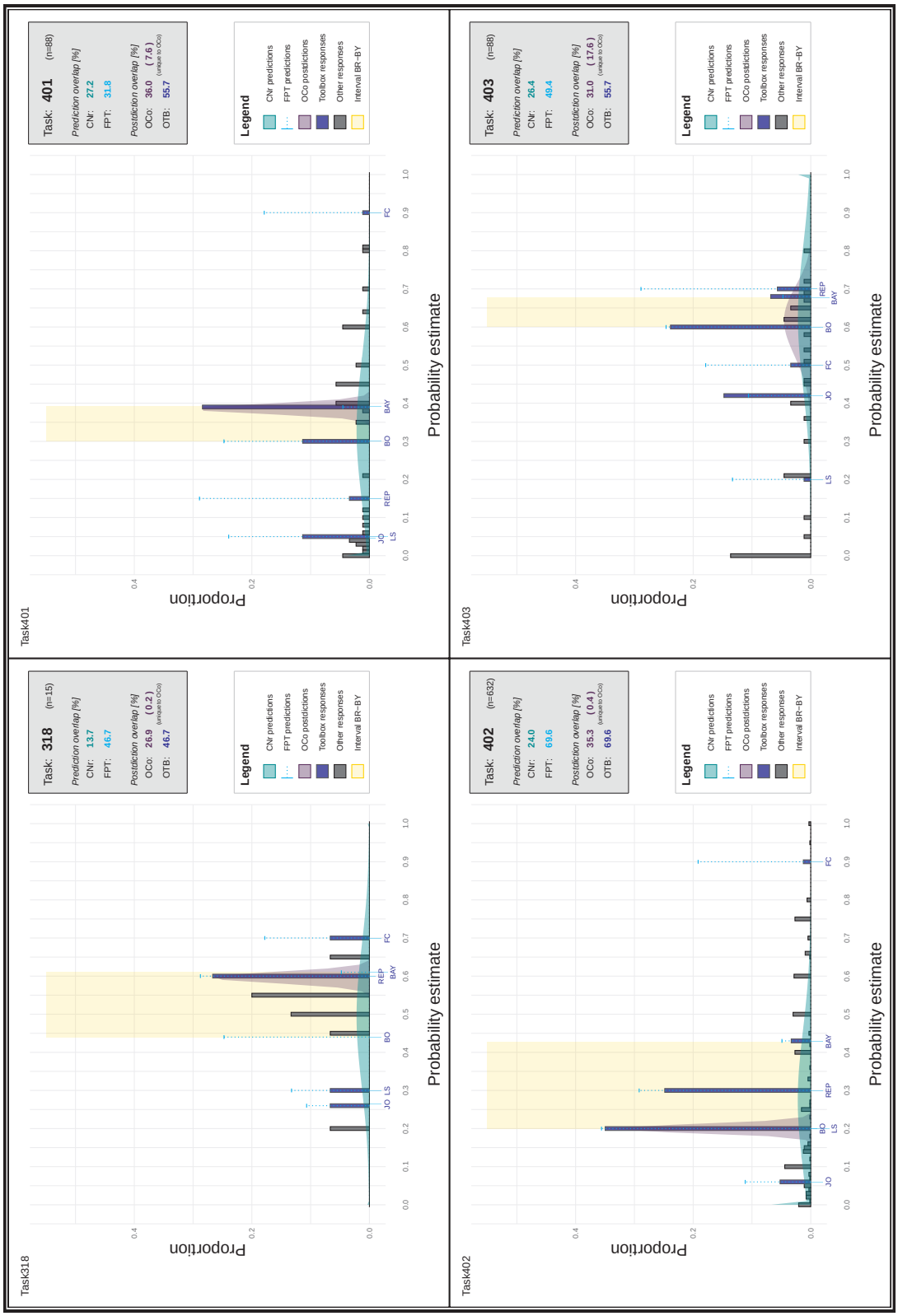

Figure S59. Comparison of toolbox and conservatism predictions across tasks (part 10/27)

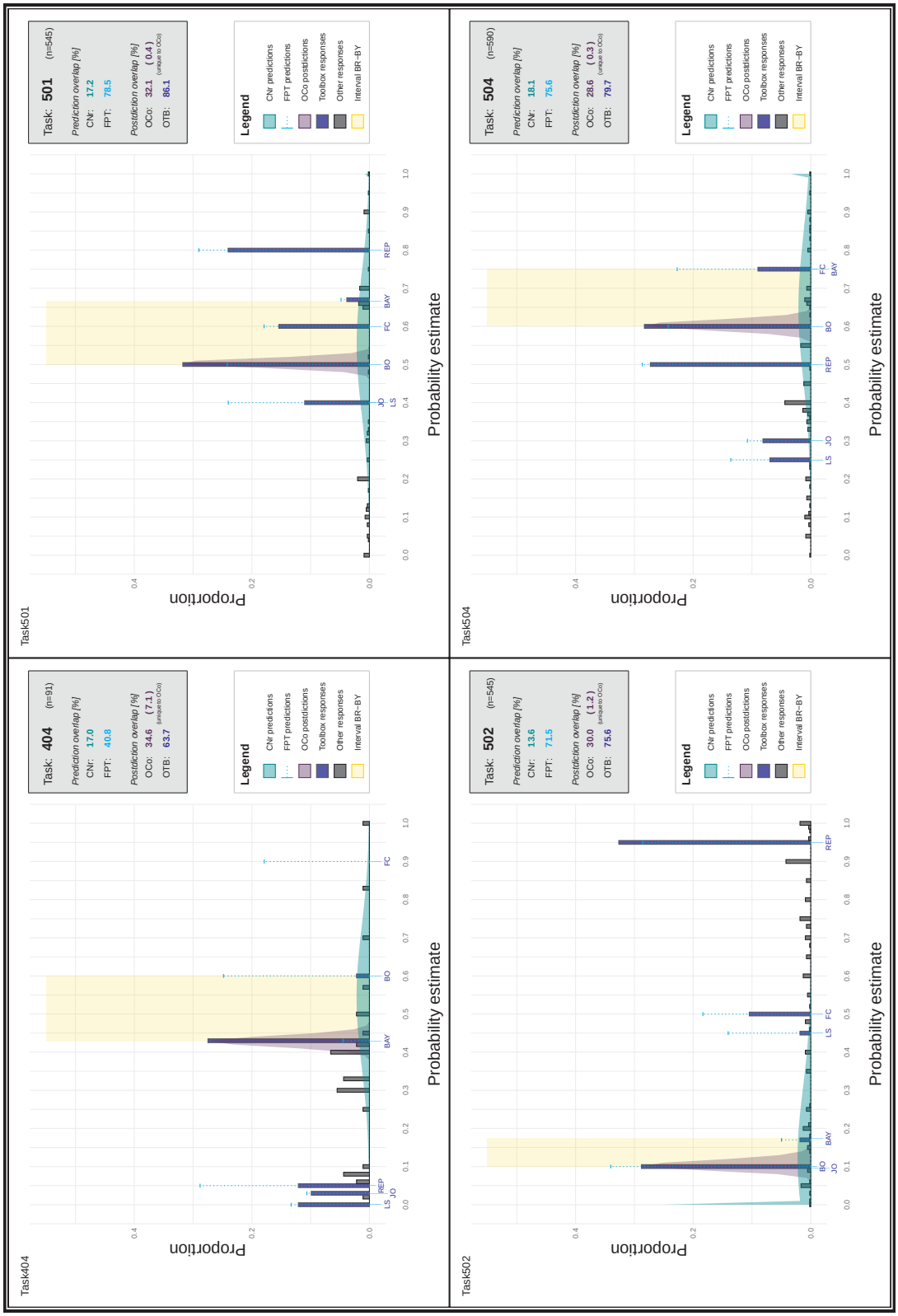

Figure S60. Comparison of toolbox and conservatism predictions across tasks (part 11/27)

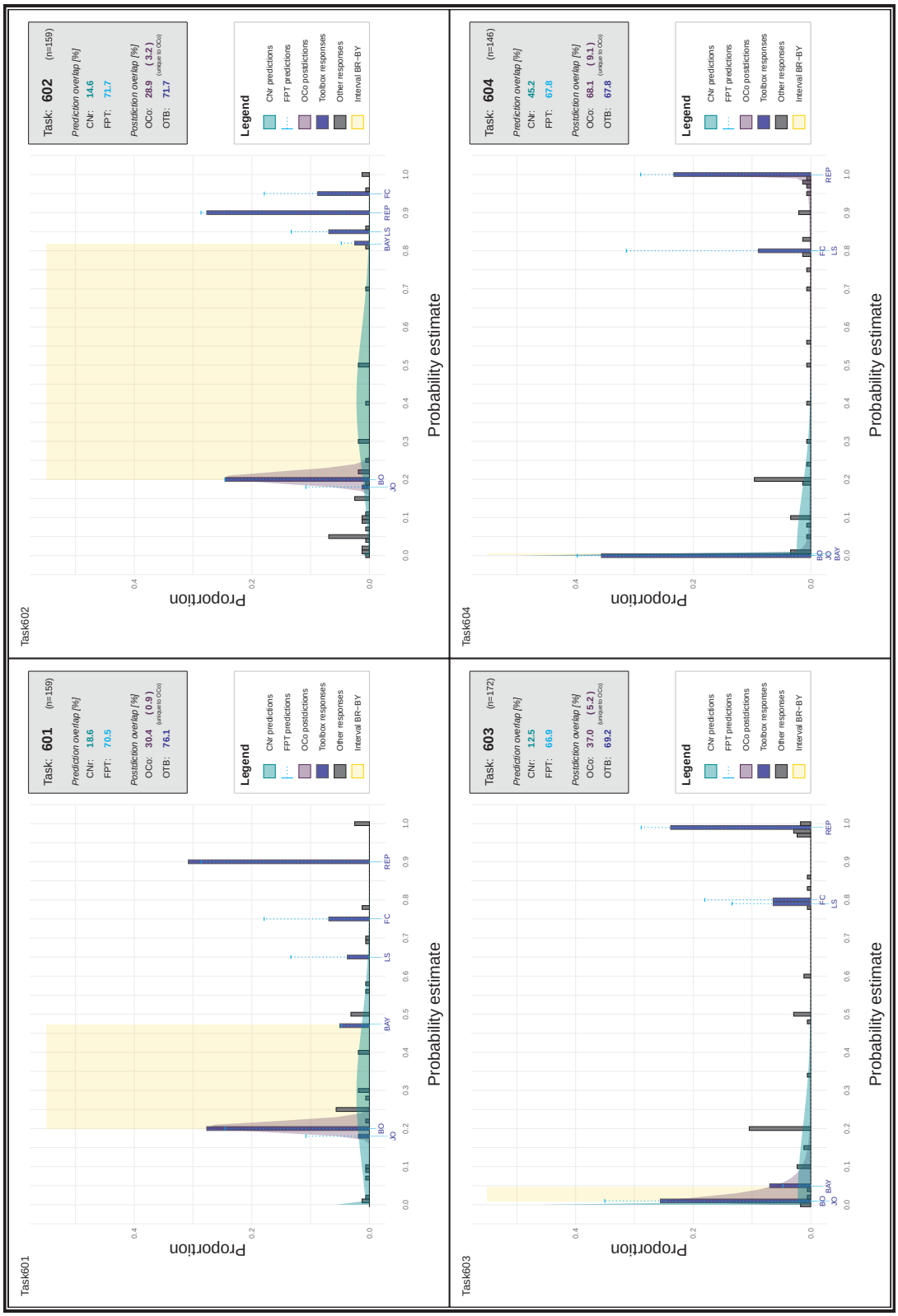

Figure S61. Comparison of toolbox and conservatism predictions across tasks (part 12/27)

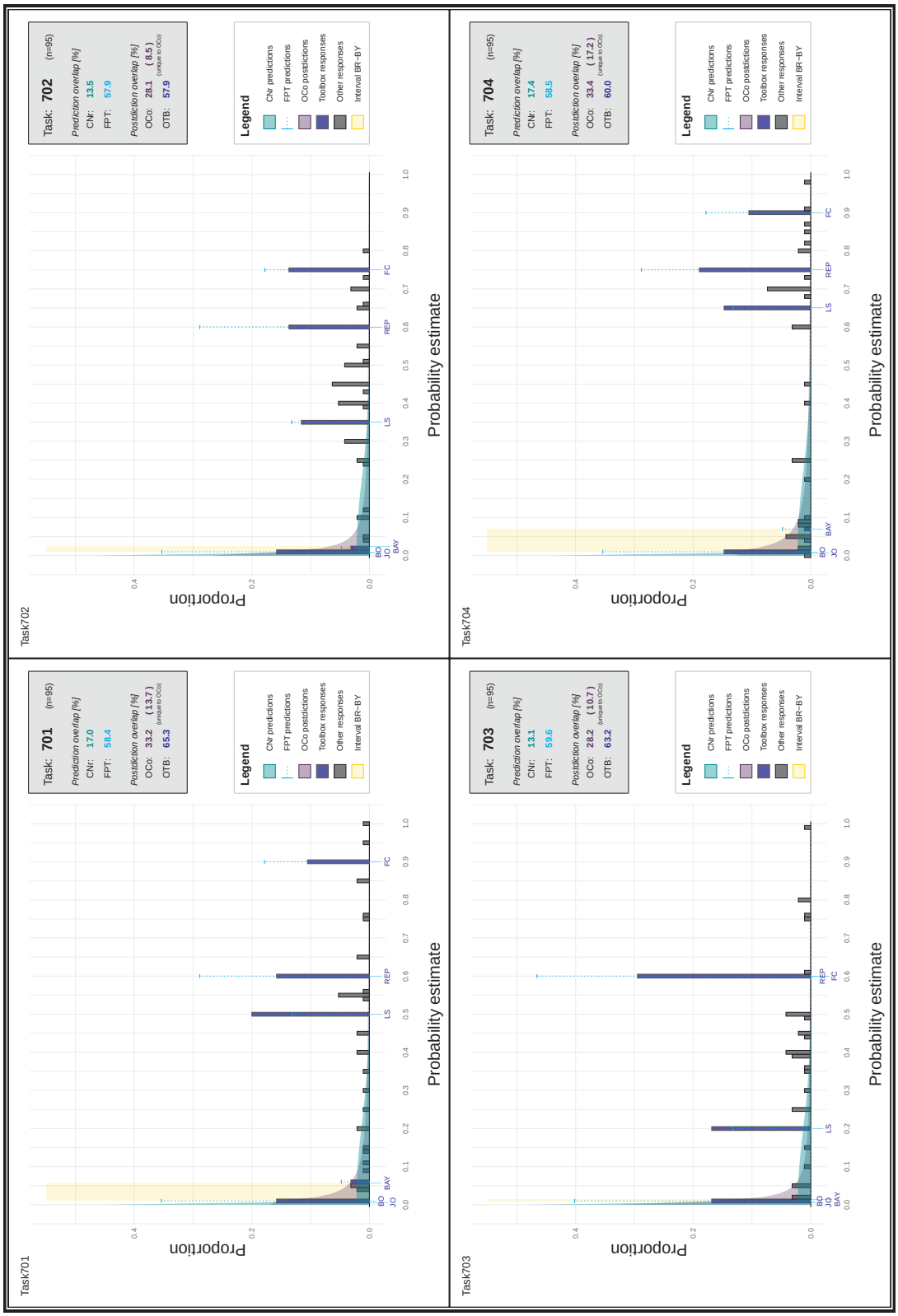

Figure S62. Comparison of toolbox and conservatism predictions across tasks (part 13/27)

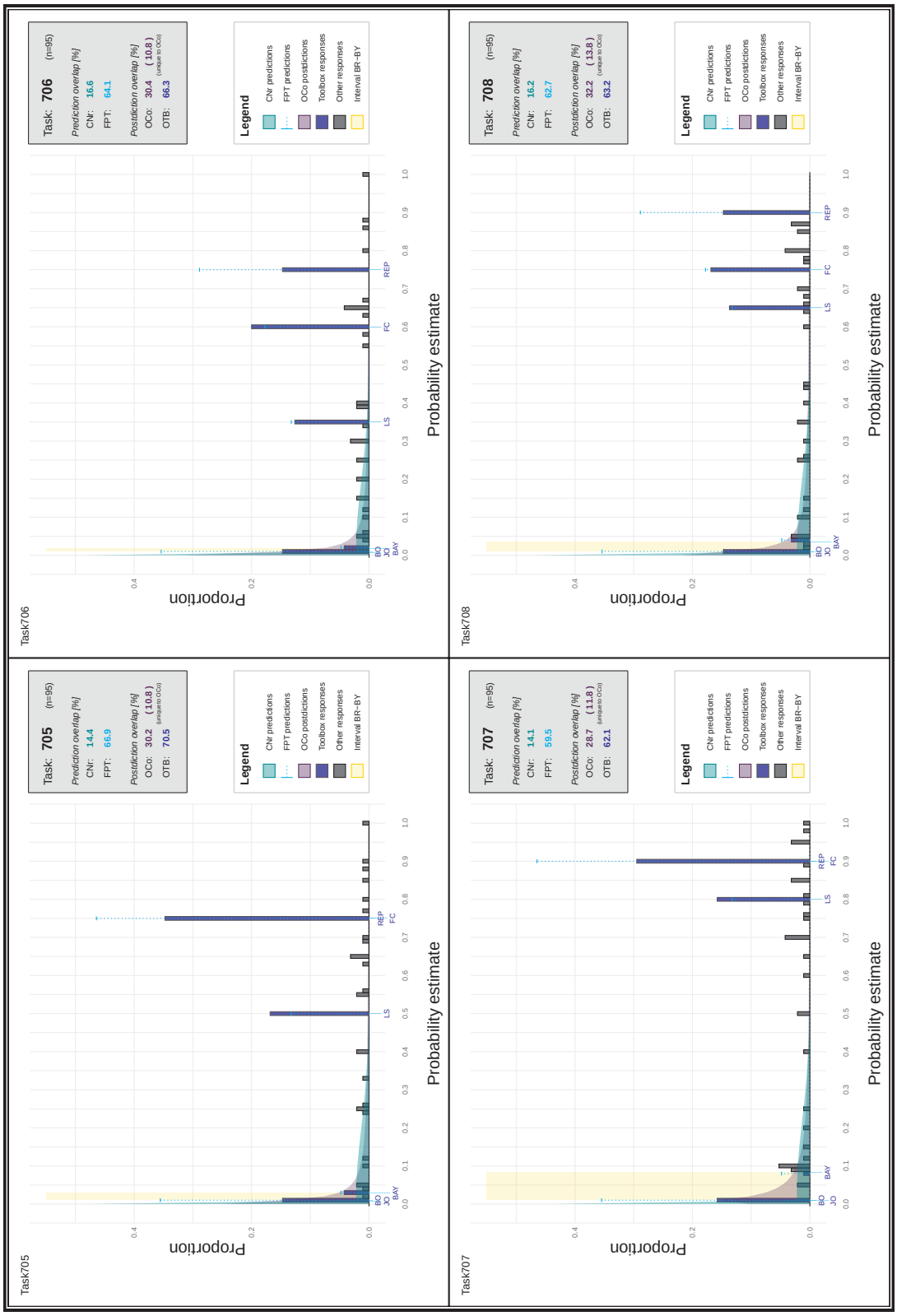

Figure S63. Comparison of toolbox and conservatism predictions across tasks (part 14/27)

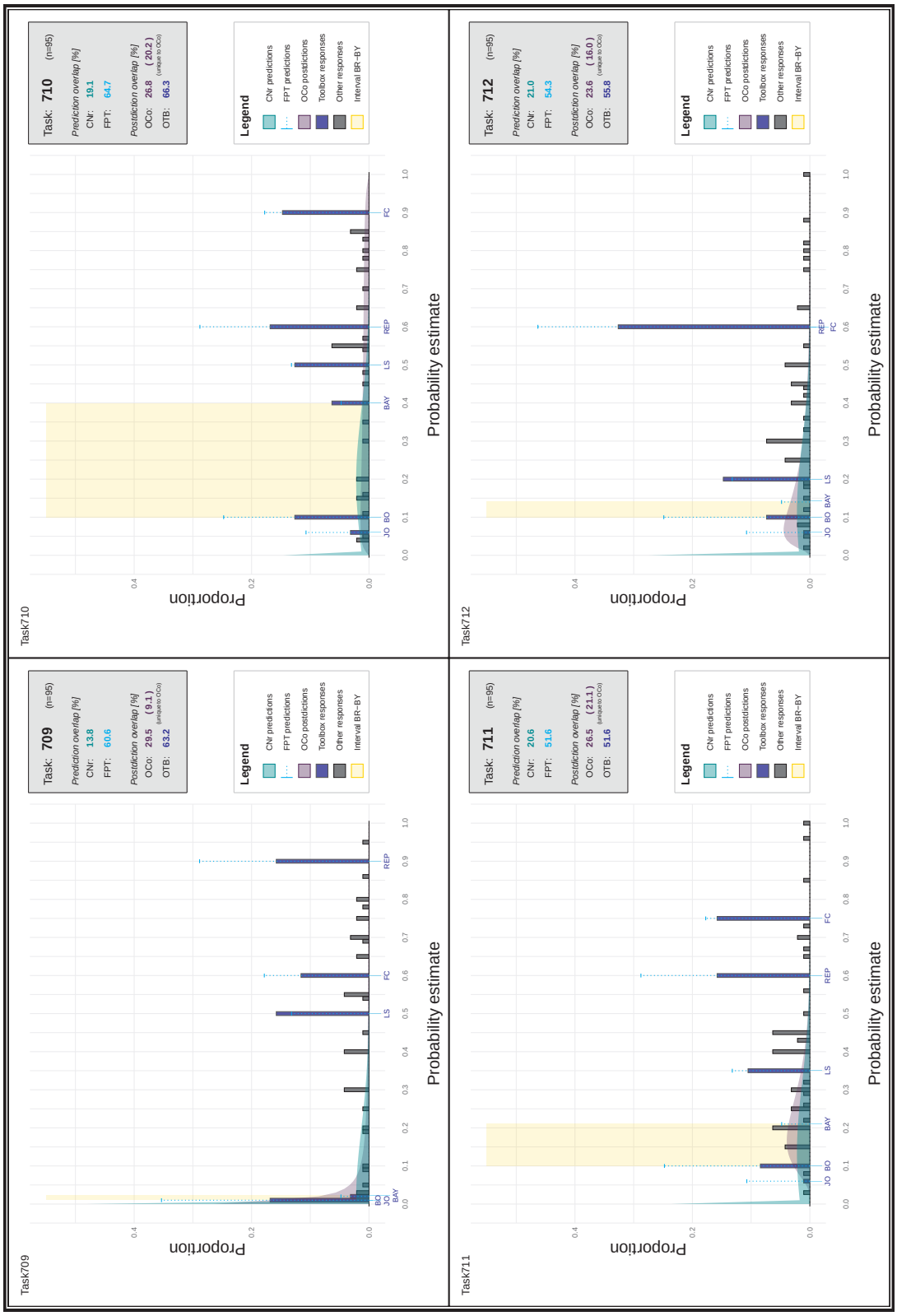

Figure S64. Comparison of toolbox and conservatism predictions across tasks (part 15/27)

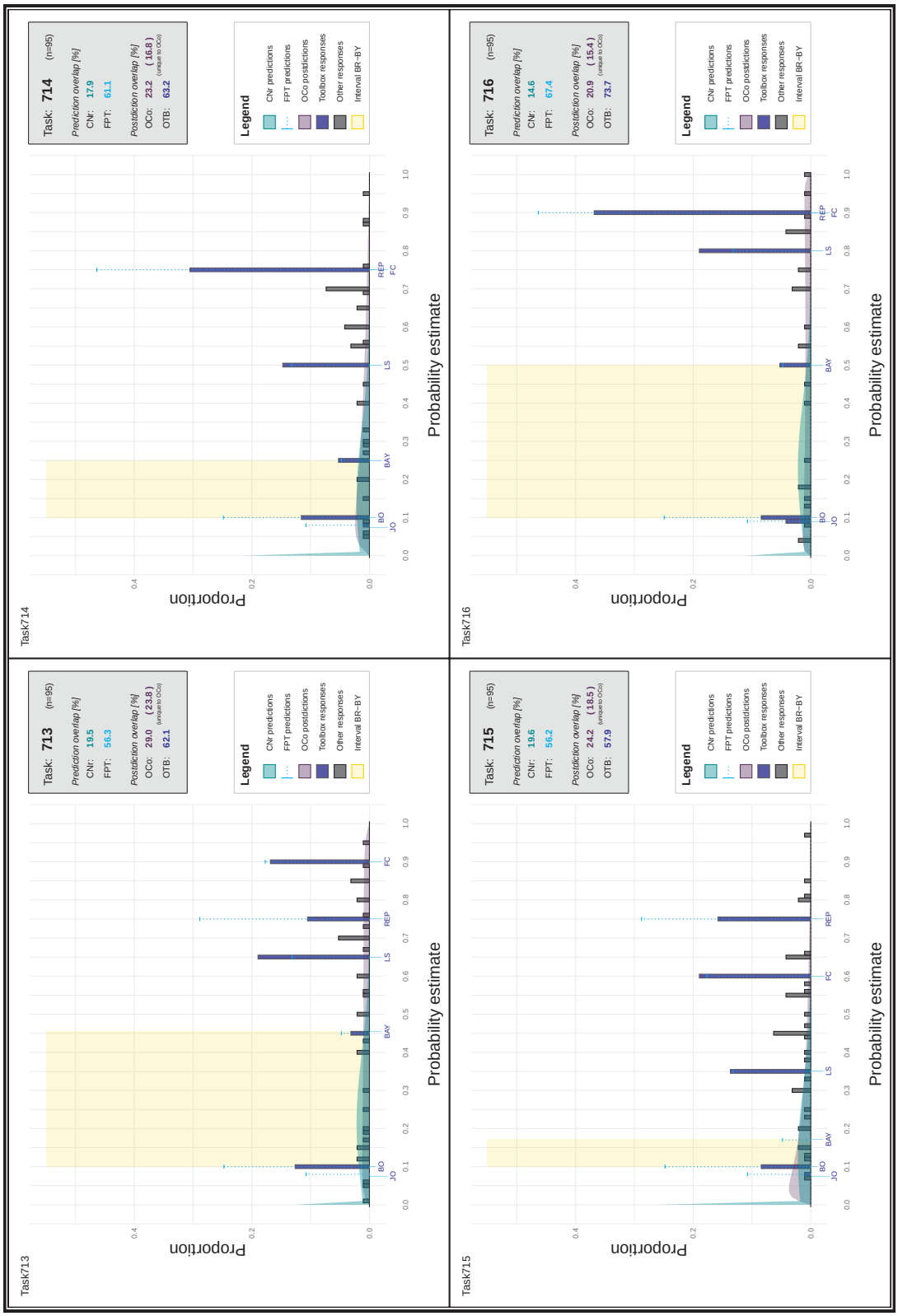

Figure S65. Comparison of toolbox and conservatism predictions across tasks (part 16/27)

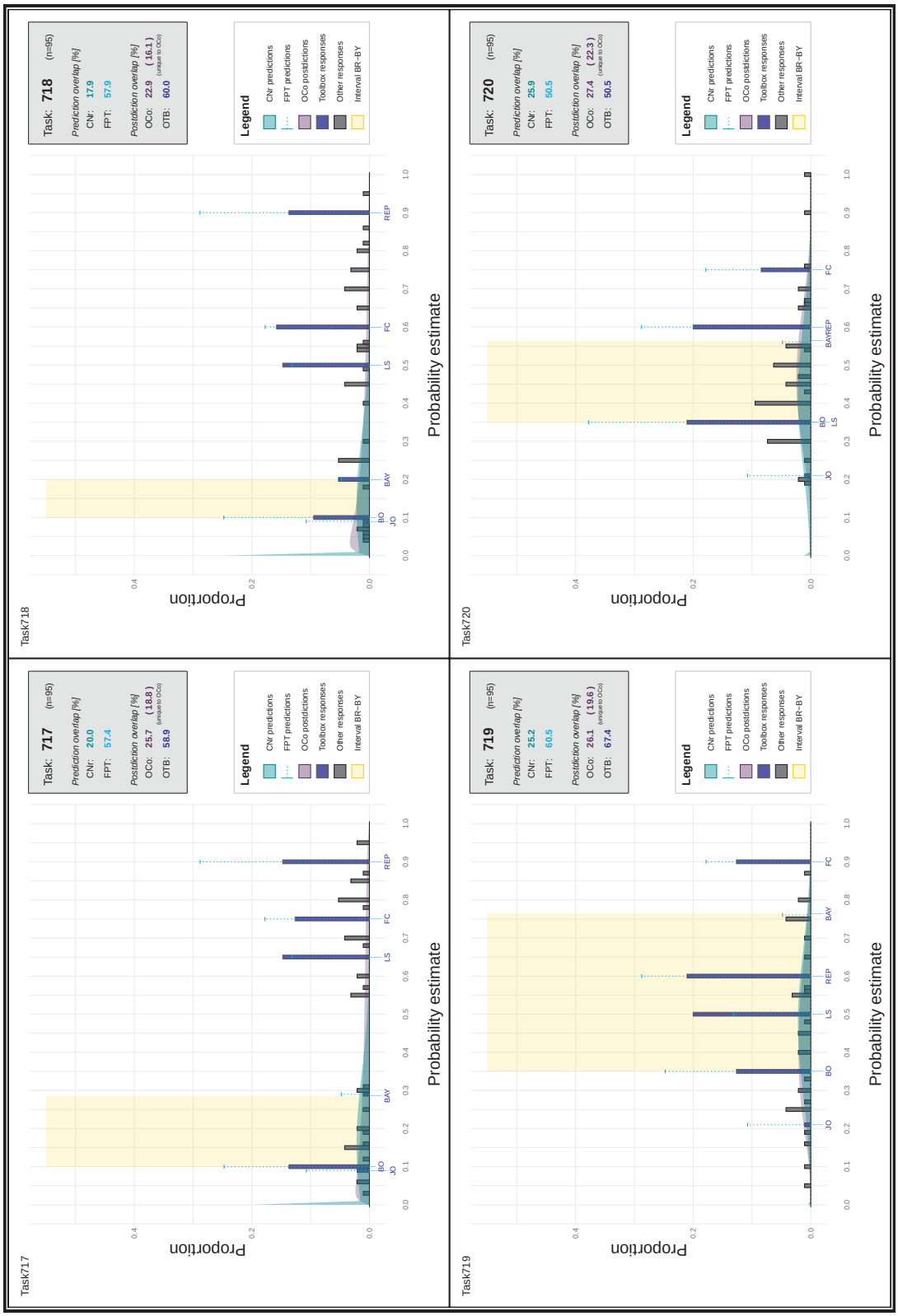

Figure S66. Comparison of toolbox and conservatism predictions across tasks (part 17/27)

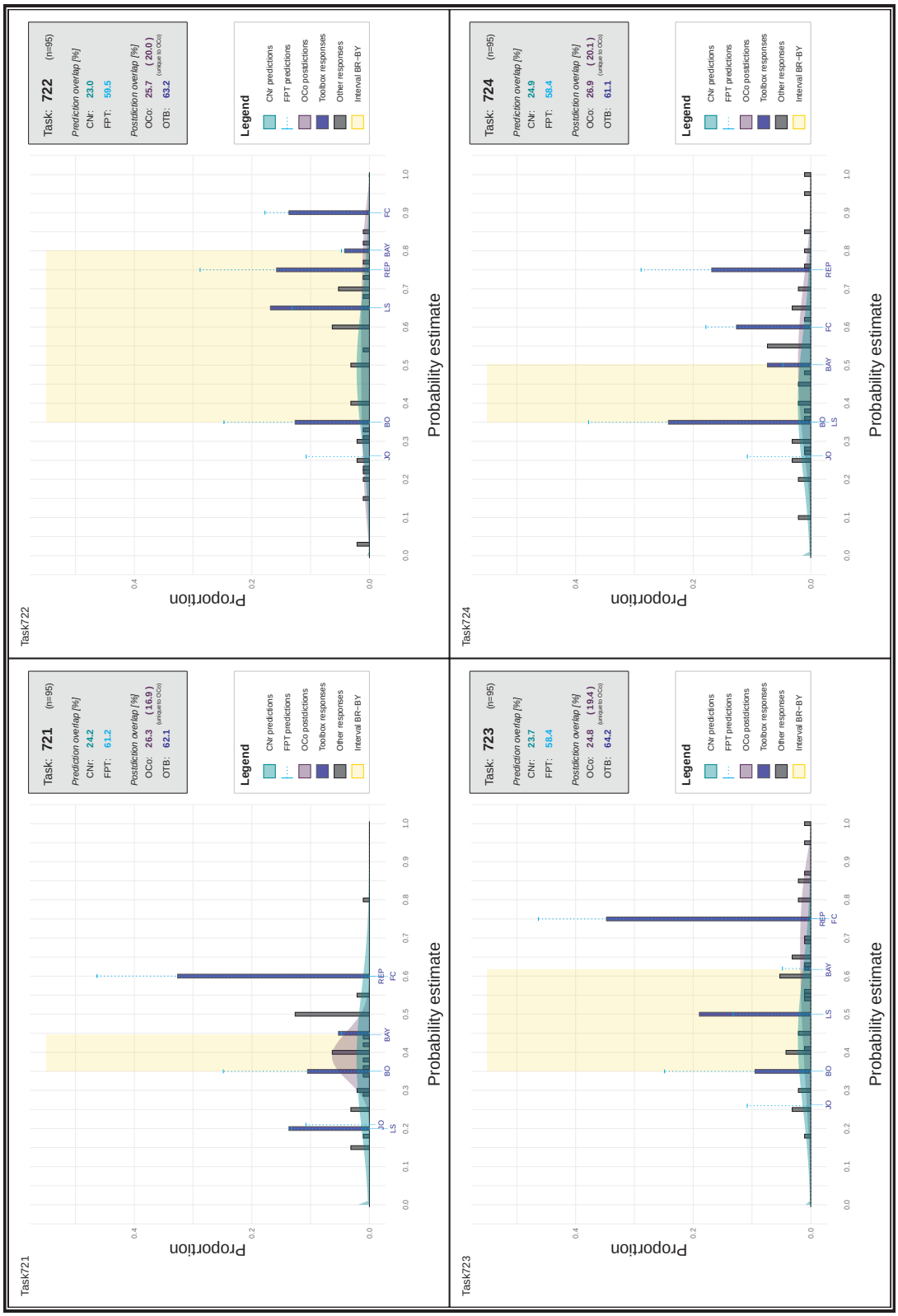

Figure S67. Comparison of toolbox and conservatism predictions across tasks (part 18/27)

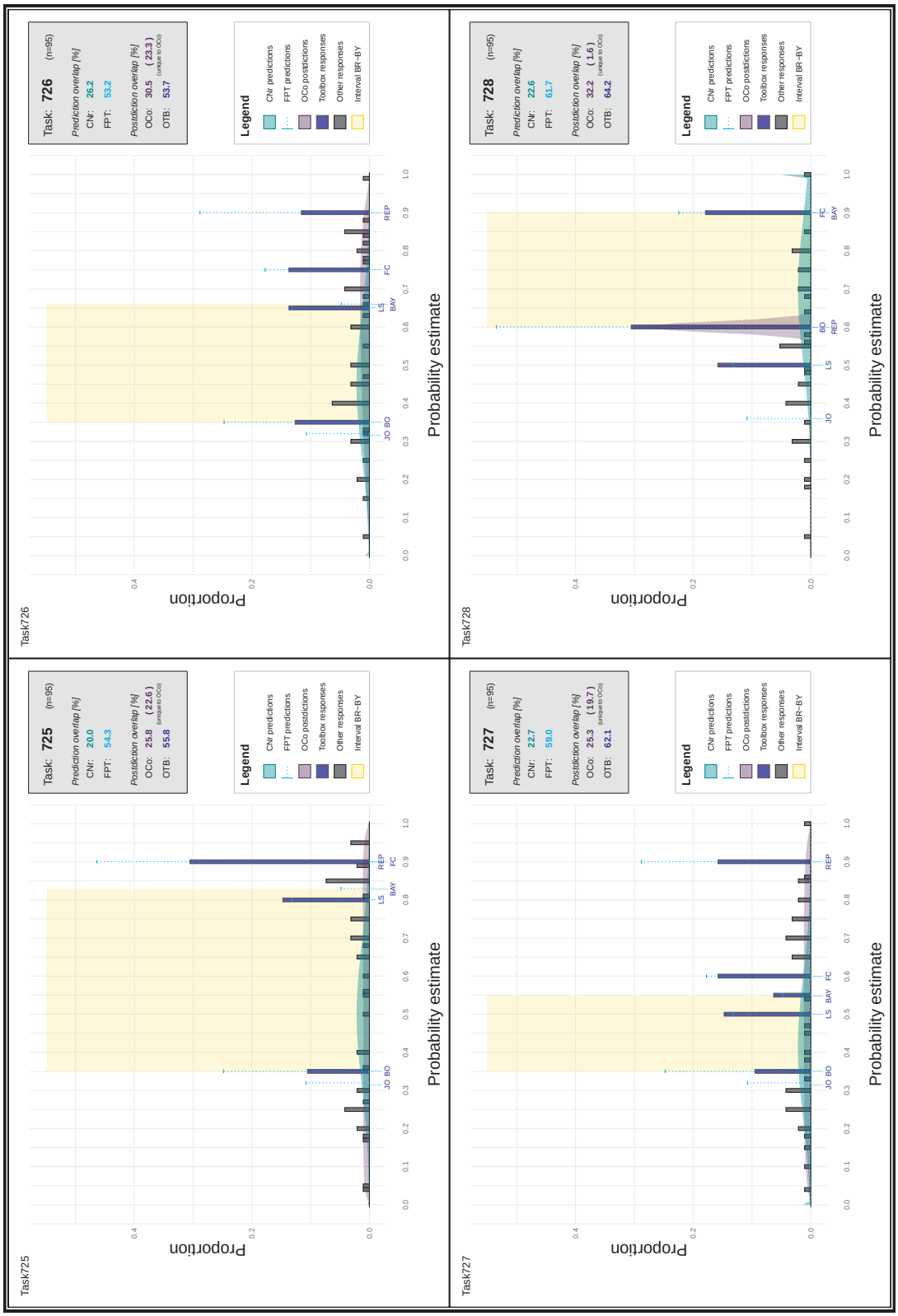

Figure S68. Comparison of toolbox and conservatism predictions across tasks (part 19/27)

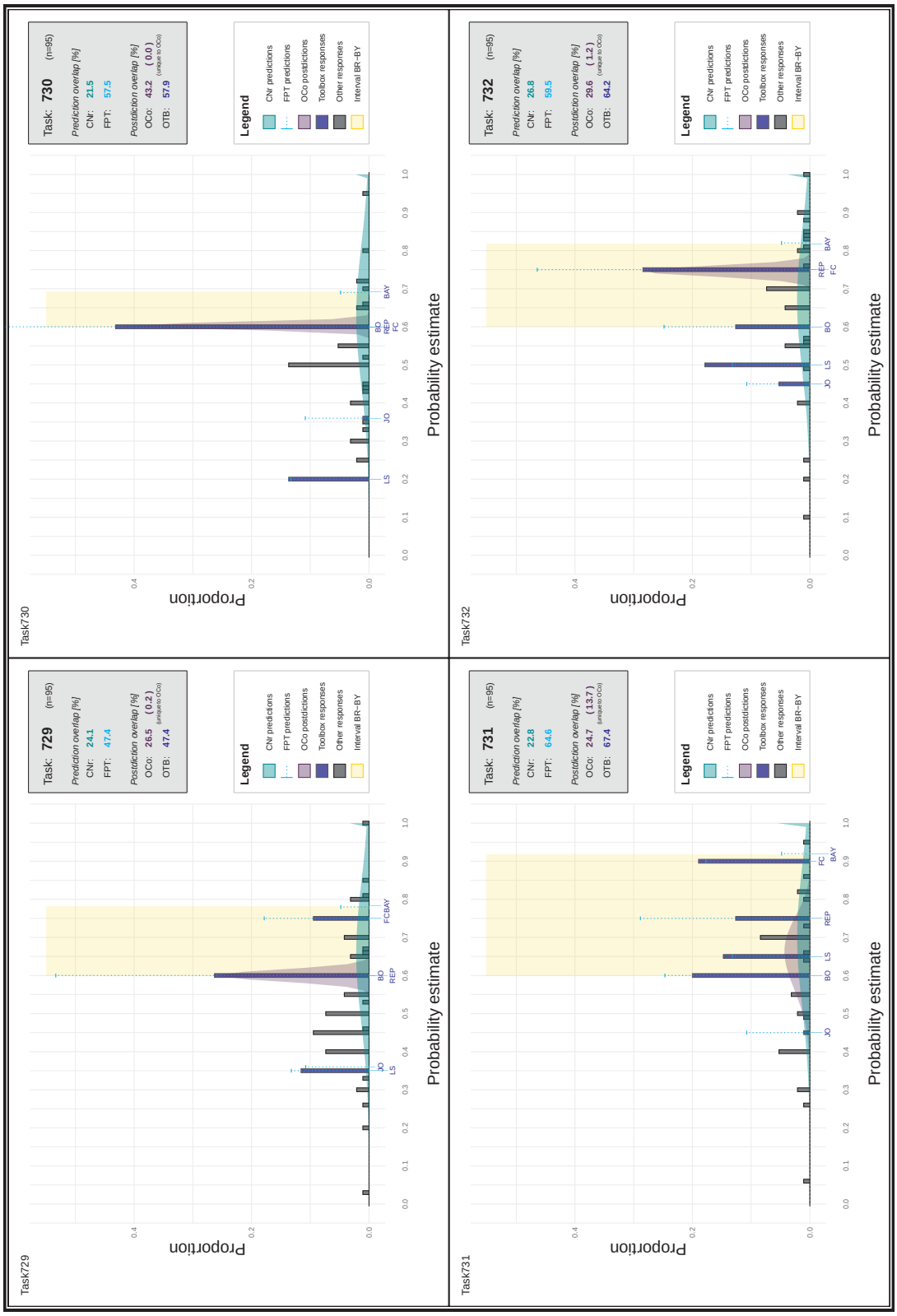

Figure S69. Comparison of toolbox and conservatism predictions across tasks (part 20/27)

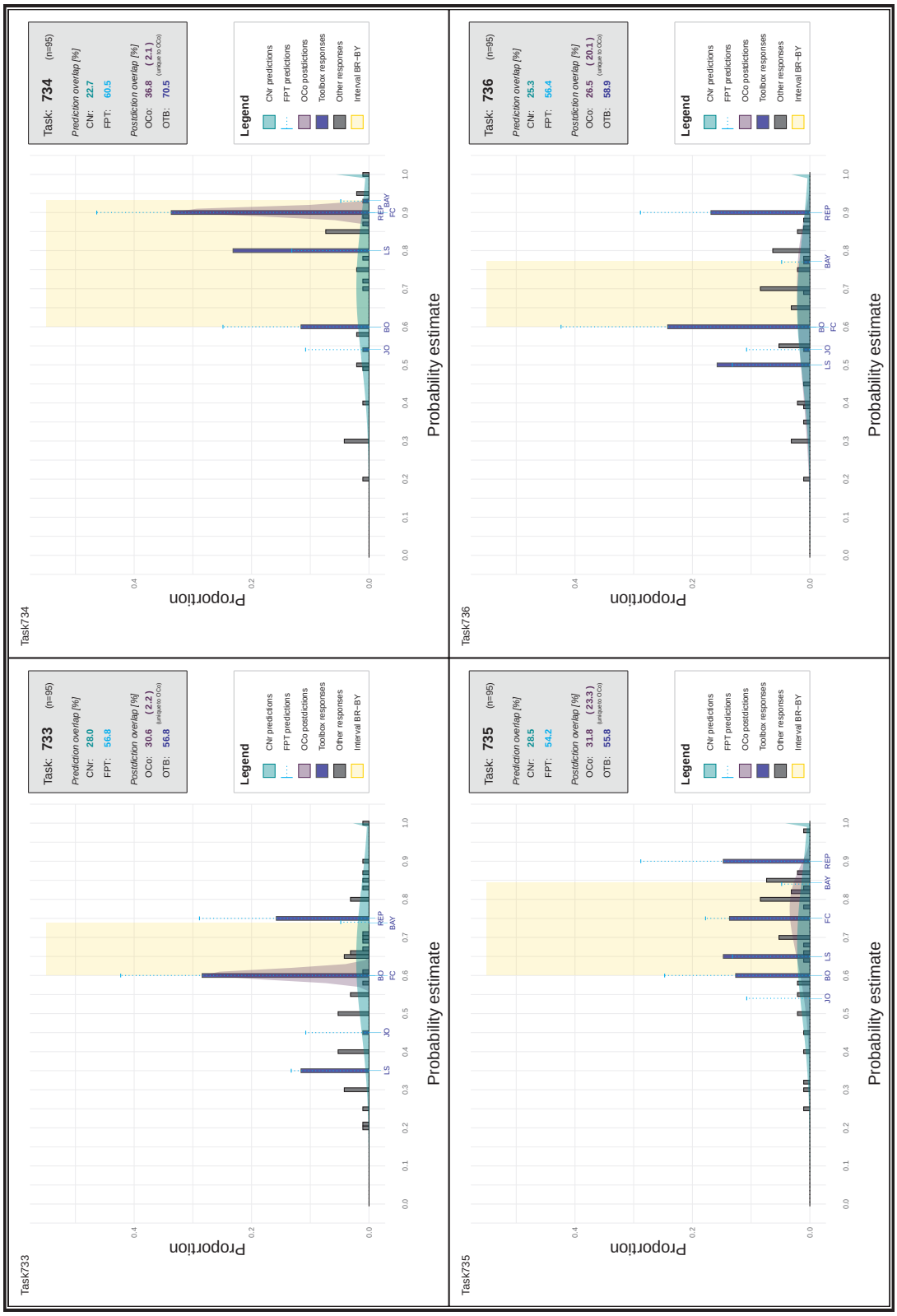

Figure S70. Comparison of toolbox and conservatism predictions across tasks (part 21/27)

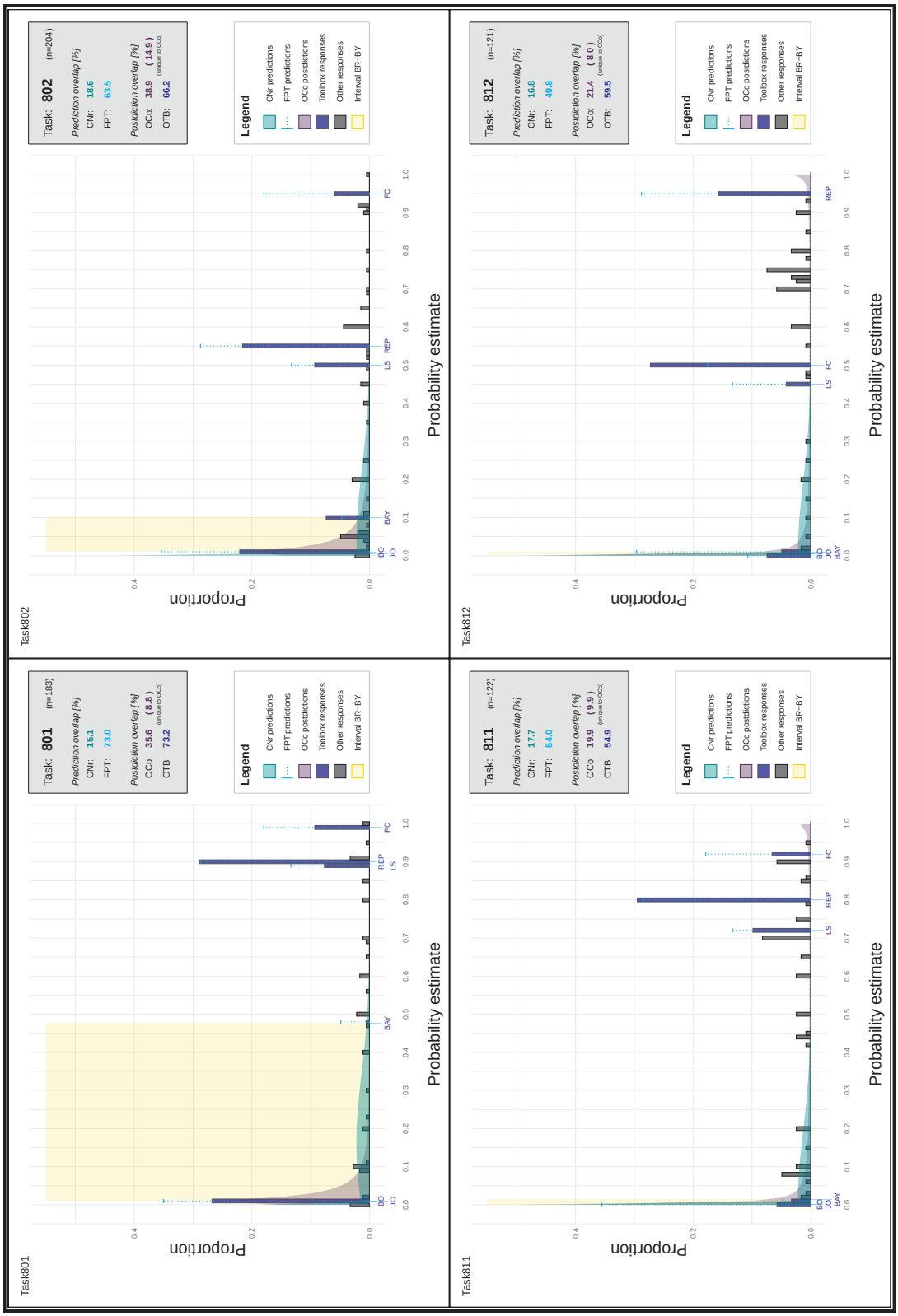

Figure S71. Comparison of toolbox and conservatism predictions across tasks (part 22/27)

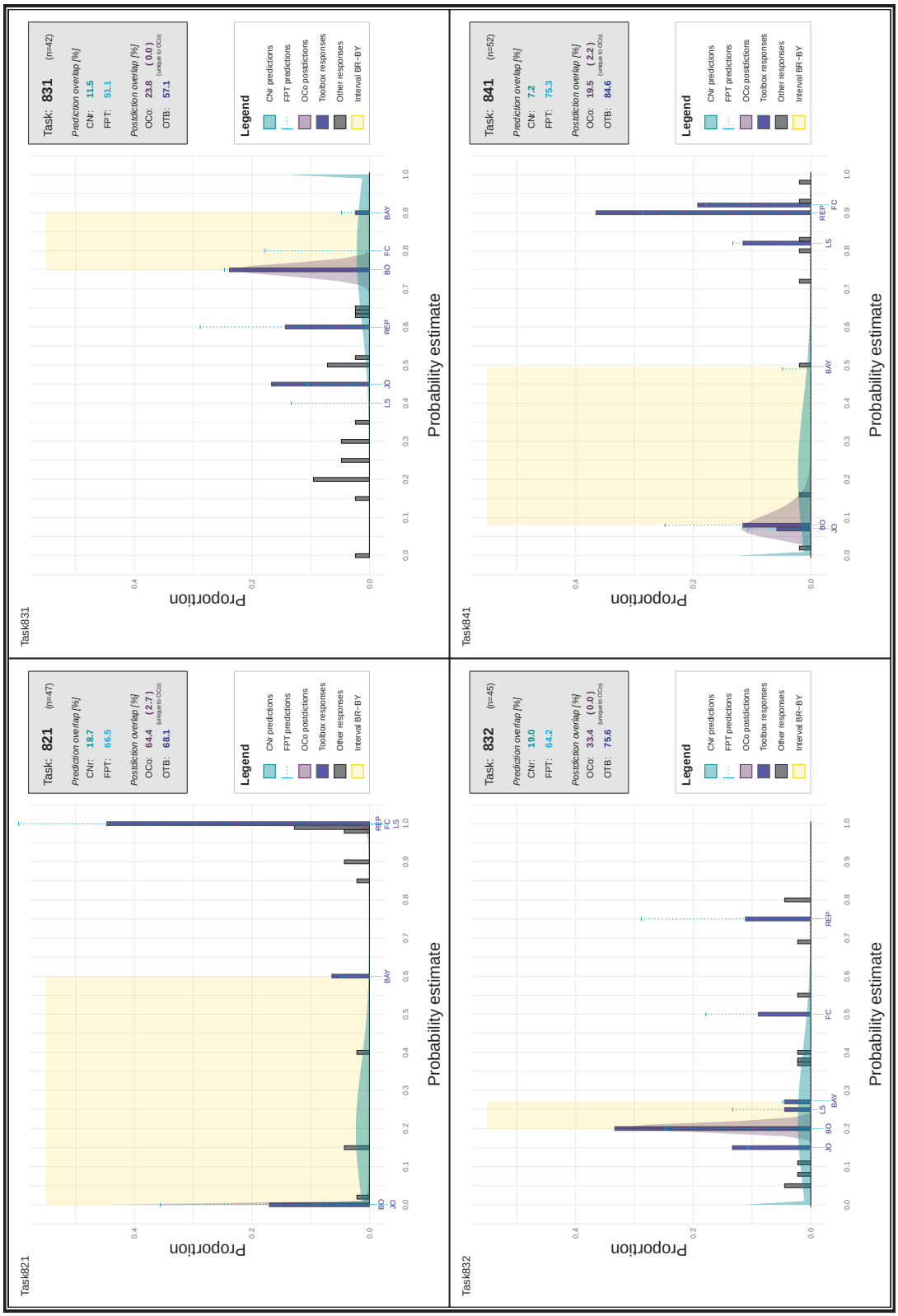

Figure S72. Comparison of toolbox and conservatism predictions across tasks (part 23/27)

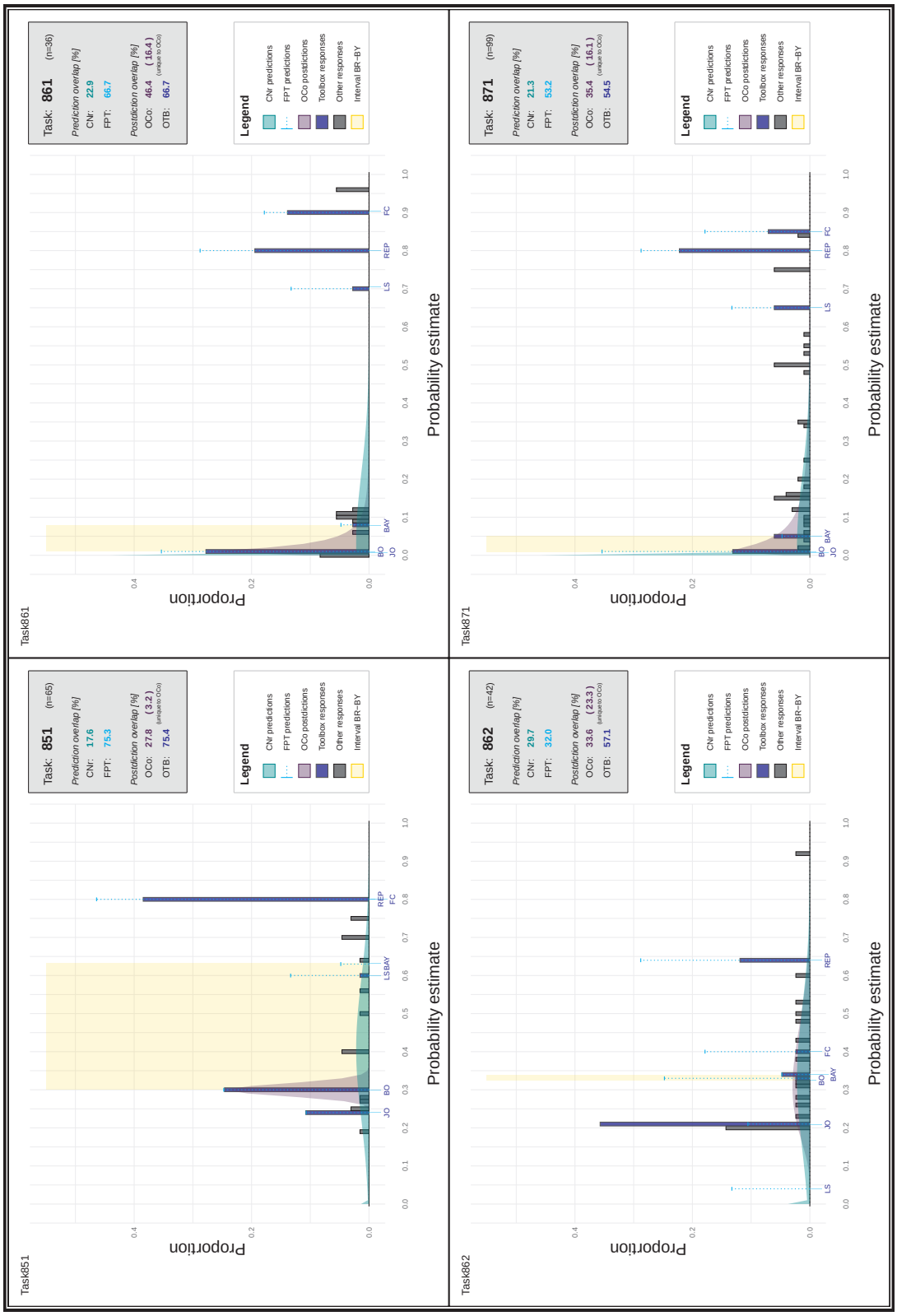

Figure S73. Comparison of toolbox and conservatism predictions across tasks (part 24/27)

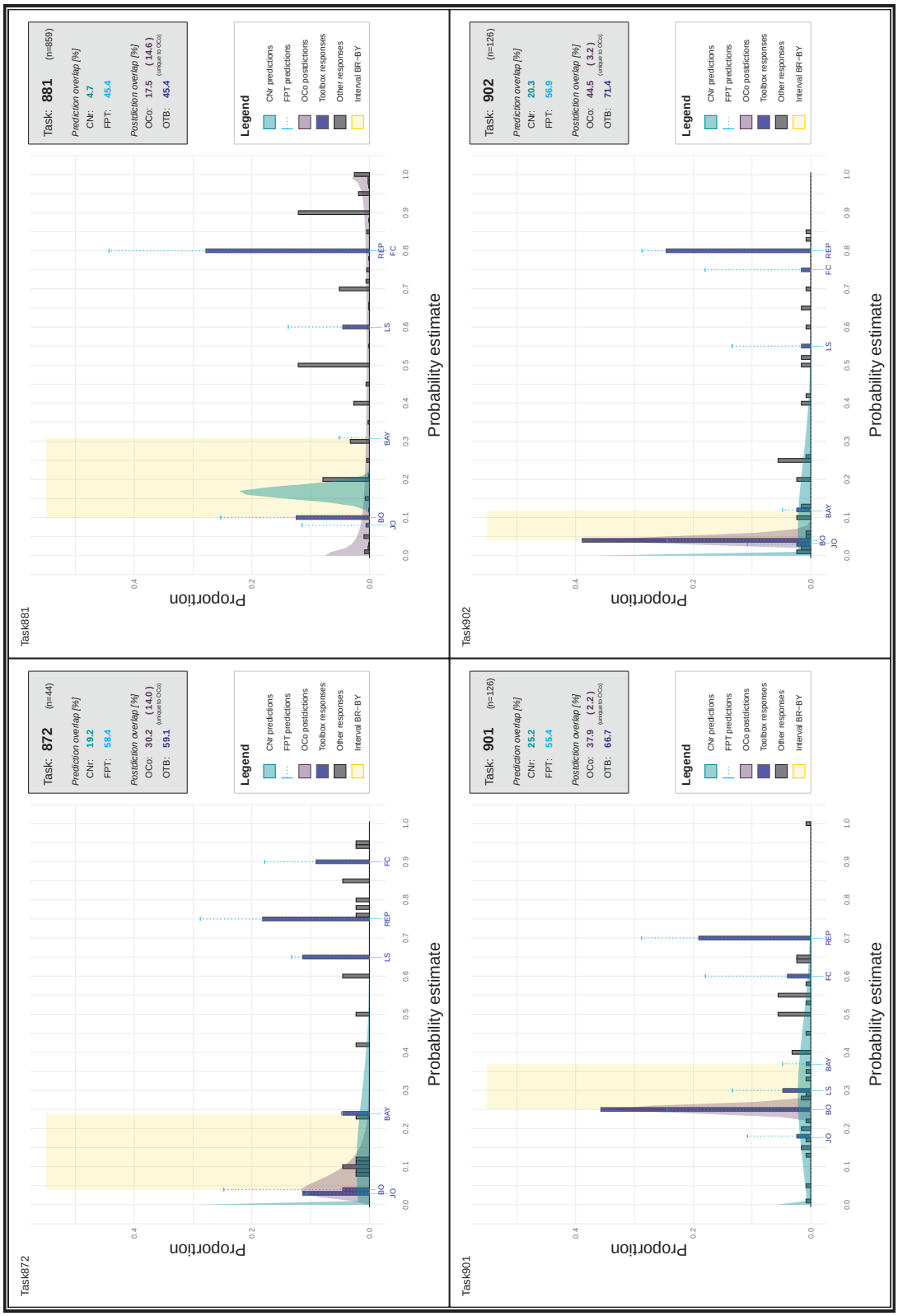

Figure S74. Comparison of toolbox and conservatism predictions across tasks (part 25/27)

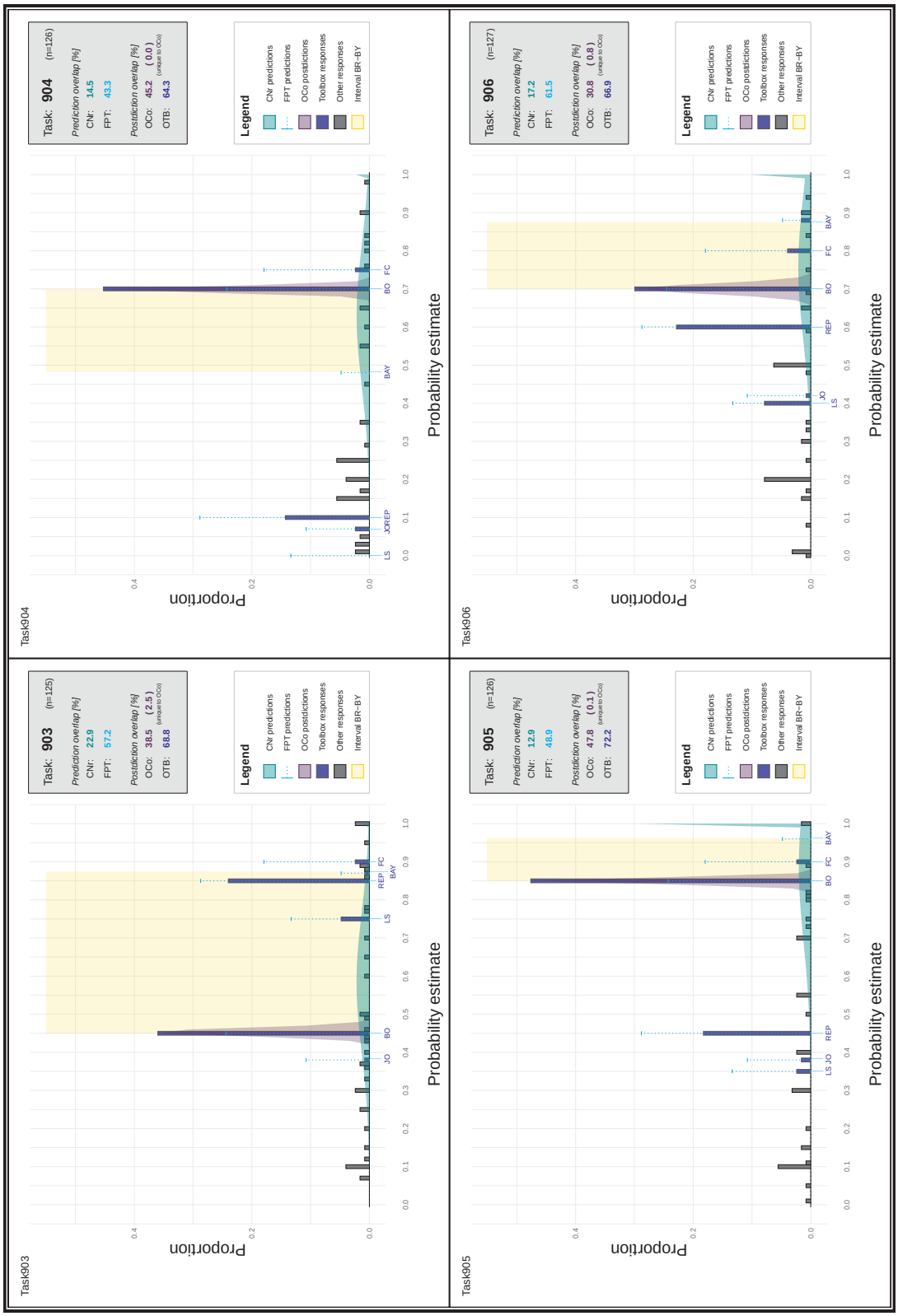

Figure S75. Comparison of toolbox and conservatism predictions across tasks (part 26/27)

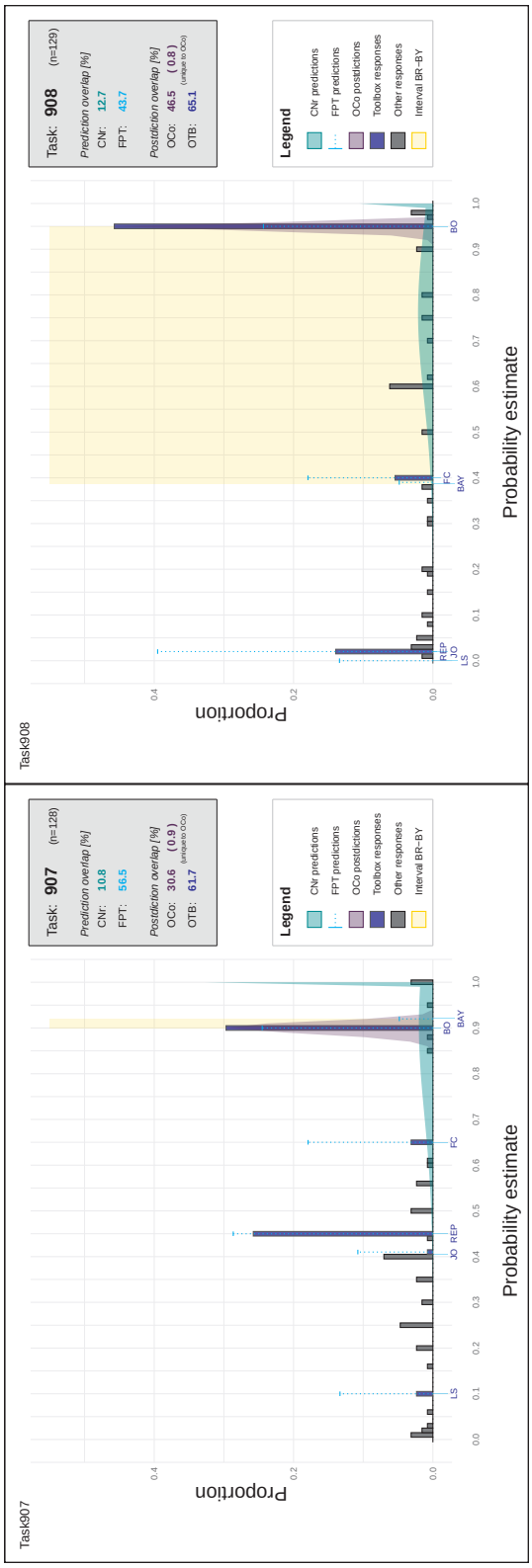

Figure S76. Comparison of toolbox and conservatism predictions across tasks (part 27/27)

7 Full list of studies

The contributing studies were published–in alphabetical order—as Bar-Hillel (1980), Binder, Krauss, and Bruckmaier (2015), Birnbaum and Mellers (1983), Chapman and Liu (2009), Cohen and Staub (2015), Dohmen, Falk, Huffman, Felix, and Sunde (2009), Galesic, Gigerenzer, and Straubinger (2009), Gigerenzer and Hoffrage (1995), Hayes, Ngo, Hawkins, and Newell (2018), Hoffrage and Gigerenzer (1998), Hoffrage, Lindsey, and Hertwig (2000), Hoffrage, Hafenbrädl, and Bouquet (2015), Juslin et al. (2011), Nadanovsky, dos Santos, Lira-Junior, and de Oliveira (2018), Pighin, Gonzalez, Savadori, and Girotto (2016), Vallée-Tourangeau, Abadie, and Vallée-Tourangeau (2015), and Weber, Binder, and Krauss (2018). We are grateful to all authors of the original studies for sharing their data, either via online repositories, full reproduction of individual data in the manuscripts, or responses to data requests. All contacted authors were very supportive and willing to share the relevant datasets. In addition, we included the data from three data collections in the context of this manuscript, named Studies 4a, 4b, and 4c.

We received additional datasets that were not included in the final set (Binder, Krauss, Bruckmaier, & Marienhagen, 2018; McNair & Feeney, 2015), as the employed tasks differed in their formal structure, either not including all pieces of information explicitly, or adding multiple pieces of diagnostic information.

The studies used in the manuscript are listed in Table S9. Tasks that used the same numbers were grouped as one task, even if the numbers were presented in different scenarios or problem texts.

Table S9

Extended sample descriptions for the empirical datasets used in the simulations. Each sample is described by the source (or Study number for our samples), the sampled population, number of participants, the maximum number of responses per participant, the total number of responses, the number and IDs of different tasks. Task IDs correspond to the numbers in Table S10.

ID	Source	Sample	N	$\frac{\text{Max responses}}{\text{Participant}}$	Responses	Tasks	Task IDs
1	Gigerenzer and Hoffrage (1995)	Students	30	15	433	15	1–15
2	Hoffrage and Gigerenzer (1998)	Physicians	48	2	96	4	201–204
3	Hoffrage et al. (2000)	Students/Prof.	66	1	66	4	201–204
4	Juslin et al. (2011)	Students	15	18	270	18	301–318
5	Hoffrage et al. (2015)	EMBA	15	2	30	4	401—404
6	Hoffrage et al. (2015)	EMBA	14	2	28	4	401–404
7	Hoffrage et al. (2015)	EMBA	10	2	20	4	401–404
8	Hoffrage et al. (2015)	EMBA	24	2	48	4	401–404
9	Hoffrage et al. (2015)	Managers	14	2	28	4	401–404
10	Hoffrage et al. (2015)	Managers	17	2	34	4	401–404
11	Hoffrage et al. (2015)	Students	37	2	74	4	401–404
12	Hoffrage et al. (2015)	Students	21	2	42	4	401–404
13	Hoffrage et al. (2015)	Students	25	2	50	4	401–404
14	Study 4a	MTurk	545	4	2180	4	402, 501–502, 504
15	Study 4b	MTurk	318	2	638	4	601–604
16	Cohen and Staub (2015)	MTurk	95	36	3420	36	701–736
17	Chapman and Liu (2009)	Students	345	1	345	2	801, 802
18	Galesic et al. (2009)	Younger adults	59	2	118	2	811–812
19	Galesic et al. (2009)	Older adults	23	2	46	2	811–812
20	Pighin et al. (2016)	MTurk	126	1	126	3	811–812, 821
21	Vallée-Tourangeau et al. (2015)	Volunteers	45	3	132	3	504, 831–832
22	Nadanovsky et al. (2018)	Dentists	52	1	52	1	841
23	Bar-Hillel (1980)	Univ. applicants	52	1	52	1	1
24	Birnbaum and Mellers (1983)	Students	65	1	65	1	851
25	Binder et al. (2015)	Students (16–18)	78	1	78	2	861–862
26	Weber et al. (2018)	Students	78	1	78	2	4, 802
27	Hayes et al. (2018), Exp. 1	MTurk	44	2	88	2	871–872
28	Hayes et al. (2018), Exp. 2	MTurk	55	1	55	1	871
29	Dohmen et al. (2009)	Representat.	859	1	859	1	881
30	Study 4c	MTurk	1013	1	1013	8	901–908

8 Full list of tasks

Table S10

List of tasks used in the manuscript: Each line details the base rate (b), the hit rate h), the false-alarm rate (f), the normative Bayesian solution (Bayes), as well as the results for joint occurrence ($b \cdot h$, JO), likelihood subtraction ($h - f$, LS), and false-alarm complement ($1 - f$, FC). Representativeness (REP) corresponds to the h column, base rate only (BO) to the b column.

ID	b	h	f	Bayes	JO	LS	FC
001	0.15	0.8	0.2	0.414	0.12	0.6	0.8
002	0.01	0.8	0.09596	0.078	0.008	0.704	0.904
003	0.0001	1	0.000001	0.99	0	1	1
004	0.0001	1	0.0019	0.05	0	0.998	0.998
005	0.02	0.95	0.005102	0.792	0.019	0.945	0.995
006	0.01	0.55	0.050505	0.099	0.006	0.499	0.949
007	0.0021	0.47619	0.005	0.167	0.001	0.471	0.995
008	0.05	0.4	0.2	0.095	0.02	0.2	0.8
009	0.03	0.9	0.4	0.065	0.027	0.5	0.6
010	0.00024	0.15	0.12	0	0	0.03	0.88
011	0.8	0.75	0.25	0.923	0.6	0.5	0.75
012	0.3	0.7	0.5	0.375	0.21	0.2	0.5
013	0.05	0.004	0.021053	0.01	0	0	0.979
014	0.00005	0.8	0.0005	0.074	0	0.8	1
015	0.36	0.75	0.2	0.678	0.27	0.55	0.8
201	0.01	0.8	0.1	0.075	0.008	0.7	0.9
202	0.003	0.5	0.03	0.048	0.002	0.47	0.97
203	0.05	0.92	0.08	0.377	0.046	0.84	0.92
204	0.0001	1	0.0005	0.167	0	1	1
301	0.5	0.6	0.02	0.968	0.3	0.58	0.98
302	0.4	0.6	0.1	0.8	0.24	0.5	0.9
303	0.01	0.9	0.3	0.029	0.009	0.6	0.7
304	0.3	0.6	0.2	0.563	0.18	0.4	0.8
305	0.1	0.9	0.2	0.333	0.09	0.7	0.8
306	0.1	0.7	0.4	0.163	0.07	0.3	0.6
307	0.33	0.9	0.1	0.816	0.297	0.8	0.9
308	0.75	0.95	0.05	0.983	0.713	0.9	0.95
309	0.41	0.8	0.05	0.917	0.328	0.75	0.95
310	0.13	0.8	0.2	0.374	0.104	0.6	0.8
311	0.17	0.6	0.48	0.204	0.102	0.12	0.52
312	0.07	0.9	0.09	0.429	0.063	0.81	0.91
313	0.074	0.9	0.36	0.167	0.067	0.54	0.64
314	0.14	0.6	0.38	0.204	0.084	0.22	0.62

continued on next page

continued from previous page

ID	b	h	f	Bayes	JO	LS	FC
315	0.24	0.6	0.24	0.441	0.144	0.36	0.76
316	0.016	0.9	0.45	0.031	0.014	0.45	0.55
317	0.059	0.5	0.4	0.073	0.03	0.1	0.6
318	0.44	0.6	0.3	0.611	0.264	0.3	0.7
401	0.3	0.15	0.1	0.391	0.045	0.05	0.9
402	0.2	0.3	0.1	0.429	0.06	0.2	0.9
403	0.6	0.7	0.5	0.677	0.42	0.2	0.5
404	0.6	0.05	0.1	0.429	0.03	0	0.9
501	0.5	0.8	0.4	0.667	0.4	0.4	0.6
502	0.1	0.95	0.5	0.174	0.095	0.45	0.5
504	0.6	0.5	0.25	0.75	0.3	0.25	0.75
601	0.2	0.9	0.25	0.474	0.18	0.65	0.75
602	0.2	0.9	0.05	0.818	0.18	0.85	0.95
603	0.01	0.99	0.2	0.048	0.01	0.79	0.8
604	0.001	0.999	0.2	0.005	0.001	0.799	0.8
701	0.01	0.6	0.1	0.057	0.006	0.5	0.9
702	0.01	0.6	0.25	0.024	0.006	0.35	0.75
703	0.01	0.6	0.4	0.015	0.006	0.2	0.6
704	0.01	0.75	0.1	0.07	0.008	0.65	0.9
705	0.01	0.75	0.25	0.029	0.008	0.5	0.75
706	0.01	0.75	0.4	0.019	0.008	0.35	0.6
707	0.01	0.9	0.1	0.083	0.009	0.8	0.9
708	0.01	0.9	0.25	0.035	0.009	0.65	0.75
709	0.01	0.9	0.4	0.022	0.009	0.5	0.6
710	0.1	0.6	0.1	0.4	0.06	0.5	0.9
711	0.1	0.6	0.25	0.211	0.06	0.35	0.75
712	0.1	0.6	0.4	0.143	0.06	0.2	0.6
713	0.1	0.75	0.1	0.455	0.075	0.65	0.9
714	0.1	0.75	0.25	0.25	0.075	0.5	0.75
715	0.1	0.75	0.4	0.172	0.075	0.35	0.6
716	0.1	0.9	0.1	0.5	0.09	0.8	0.9
717	0.1	0.9	0.25	0.286	0.09	0.65	0.75
718	0.1	0.9	0.4	0.2	0.09	0.5	0.6
719	0.35	0.6	0.1	0.764	0.21	0.5	0.9
720	0.35	0.6	0.25	0.564	0.21	0.35	0.75
721	0.35	0.6	0.4	0.447	0.21	0.2	0.6
722	0.35	0.75	0.1	0.802	0.263	0.65	0.9
723	0.35	0.75	0.25	0.618	0.263	0.5	0.75
724	0.35	0.75	0.4	0.502	0.263	0.35	0.6
725	0.35	0.9	0.1	0.829	0.315	0.8	0.9

continued on next page

continued from previous page

ID	b	h	f	Bayes	JO	LS	FC
726	0.35	0.9	0.25	0.66	0.315	0.65	0.75
727	0.35	0.9	0.4	0.548	0.315	0.5	0.6
728	0.6	0.6	0.1	0.9	0.36	0.5	0.9
729	0.6	0.6	0.25	0.783	0.36	0.35	0.75
730	0.6	0.6	0.4	0.692	0.36	0.2	0.6
731	0.6	0.75	0.1	0.918	0.45	0.65	0.9
732	0.6	0.75	0.25	0.818	0.45	0.5	0.75
733	0.6	0.75	0.4	0.738	0.45	0.35	0.6
734	0.6	0.9	0.1	0.931	0.54	0.8	0.9
735	0.6	0.9	0.25	0.844	0.54	0.65	0.75
736	0.6	0.9	0.4	0.771	0.54	0.5	0.6
801	0.01	0.9	0.01	0.476	0.009	0.89	0.99
802	0.01	0.55	0.05	0.1	0.006	0.5	0.95
811	0.0015	0.8	0.08	0.015	0.001	0.72	0.92
812	0.005	0.95	0.5	0.009	0.005	0.45	0.5
821	0.0015	1	0.001	0.6	0.002	0.999	0.999
831	0.75	0.6	0.2	0.9	0.45	0.4	0.8
832	0.2	0.75	0.5	0.273	0.15	0.25	0.5
841	0.08	0.9	0.08	0.495	0.072	0.82	0.92
851	0.3	0.8	0.2	0.632	0.24	0.6	0.8
861	0.01	0.8	0.096	0.078	0.008	0.704	0.904
862	0.325	0.64	0.6	0.339	0.208	0.04	0.4
871	0.01	0.8	0.15	0.051	0.008	0.65	0.85
872	0.04	0.75	0.1	0.238	0.03	0.65	0.9
881	0.1	0.8	0.2	0.308	0.08	0.6	0.8
901	0.25	0.7	0.4	0.368	0.175	0.3	0.6
902	0.04	0.8	0.25	0.118	0.032	0.55	0.75
903	0.45	0.85	0.1	0.874	0.383	0.75	0.9
904	0.7	0.1	0.25	0.483	0.07	0	0.75
905	0.85	0.45	0.1	0.962	0.383	0.35	0.9
906	0.7	0.6	0.2	0.875	0.42	0.4	0.8
907	0.9	0.45	0.35	0.92	0.405	0.1	0.65
908	0.95	0.02	0.6	0.388	0.019	0	0.4

9 Full list of strategies generated by researchers and participants

Table S11

Full list of rules found in the literature across 106 Bayesian reasoning tasks, ordered by how well they predict individual responses (median across tasks: MEDIAN). Also shown is the percentage of responses across all individuals and tasks (AVERAGE), and the potential of the free parameters of the weighting-and-adding rule to mimic a rule (MIMIC). In addition to WA_e and WA_j, the list includes WA_n, a normative WA-model, for which parameters were estimated after replacing empirical estimates in the full dataset by Bayesian posteriors. Sources are C (Cohen & Staub, 2015); G: (Gigerenzer & Hoffrage, 1995); J: (Juslin et al., 2009); K: (McKenzie, 1994); Ma: (Macchi, 2000); Me: (Mellers & McGraw, 1999); Z: (Zhu & Gigerenzer, 2006); S4: Studies 4a, and 4b in this manuscript.

Rank	Rule	Abbrev.	Source	Median	Average	Mimic
1	h	REP	C,G,K,Me,S4,Z	24.0	23.8	yes
2	b	BO	C,G,Ma,Me,S4,Z	11.6	20.5	yes
3	$1-f$	FC	C,G,Ma,S4	11.1	14.7	yes
4	$h-f$	LS	G,Ma,S4	8.8	11.1	yes
5	$b \cdot h$	JO	G,Me,S4	6.7	8.8	no
6	0.5	50%	Me,S4	3.1	9.0	yes
7	$b \cdot (h+f)$	—	S4	2.5	8.2	no
8	$\frac{b \cdot h}{b \cdot h + (1-b) \cdot f}$	Bayes	All	2.4	3.9	no
9	f	FO	C,Me,S4	2.4	5.8	yes
10	$\frac{h+f}{2}$	—	S4	1.1	6.3	yes
11	$0.5 \cdot b$	50%	S4	1.1	5.6	yes
12	$\frac{b \cdot h}{(1-b) \cdot f}$	—	S4	1.1	2.4	no
13	$\frac{h}{h+f}$	—	K,Ma	1.0	8.9	no
14	$b \cdot h + (1-b) \cdot f$	—	Me,Z	0.8	2.7	no
15	$h-b$	—	G	0.6	3.2	yes
16	$0.5 \cdot (h + \frac{h}{h+f})$	—	K	0	7.9	no
17	$1-(b \cdot h + (1-b) \cdot f)$	—	G,Ma	0	5.8	no
18	$\frac{b}{b+h}$	—	S4	0	5.4	no
19	$b \cdot (h-f)$	—	Ma,S4	0	4.2	no
20	$b+h-f$	—	S4	0	3.7	yes
21	$b+f$	—	S4	0	3.4	yes
22	$\frac{b+h+f}{3}$	—	S4	0	3.1	yes
23	$\frac{b}{h}$	—	S4	0	3.1	no
24	$0.316b + 0.435h - 0.128f + 0.148$	WA_e	J	0	3.0	(yes)

continued on next page

continued from previous page

Rank	Rule	Abbrev.	Source	Median	Average	Mimic
25	$\frac{f}{h+f}$	—	S4	0	2.9	no
26	$h+f$	—	Ma	0	2.8	yes
27	$f \cdot (1-b)$	—	Ma	0	2.7	no
28	$b+h$	—	Ma,S4	0	2.7	yes
29	$b \cdot h+f$	—	Ma	0	2.5	no
30	$b-f$	—	Ma,S4	0	2.2	yes
31	$\frac{b+h}{2}$	—	K,S4	0	2.0	yes
32	$h \cdot f$	—	S4	0	1.9	no
33	$b \cdot (b \cdot h+(1-b) \cdot f)$	—	G	0	1.8	no
34	$b \cdot h-f$	—	S4	0	1.6	no
35	$b \cdot h \cdot (b \cdot h+(1-b) \cdot f)$	—	G	0	1.6	no
36	$\frac{h}{f}$	—	S4	0	1.6	no
37	$\frac{h+f}{f}$	—	S4	0	1.6	no
38	$0.092 \cdot b+0.908 \cdot \frac{b \cdot h}{b \cdot h+(1-b) \cdot f}$	CON	—	0	1.5	no
39	$h \cdot (b \cdot h+(1-b) \cdot f)$	—	G	0	1.4	no
40	$0.680 \cdot b+0.320 \cdot \frac{b \cdot h}{b \cdot h+(1-b) \cdot f}$	CON_r	—	0	1.3	no
41	$1-\frac{h \cdot b}{f}$	—	S4	0	1.1	no
42	$1.176b+0.424h-0.736f-0.008$	WA_n	J	0	1.0	(yes)
43	$\frac{b}{b \cdot h+(1-b) \cdot f}$	—	Z	0	0.9	no
44	$0.61b+0.46h-0.4f+0.18$	WA_j	J	0	0.9	(yes)
45	$b+h \cdot f$	—	S4	0	0.8	no

10 References

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*(3), 211–233. doi: 10.1016/0001-6918(80)90046-3

Binder, K., Krauss, S., & Bruckmaier, G. (2015). Effects of visualizing statistical information– an empirical study on tree diagrams and 2× 2 tables. *Frontiers in Psychology*, *6*, 1186. doi: 10.3389/fpsyg.2015.01186

Binder, K., Krauss, S., Bruckmaier, G., & Marienhagen, J. (2018). Visualizing the Bayesian 2-test case: The effect of tree diagrams on medical decision making. *PloS one*, *13*(3), e0195029. doi: 10.1371/journal.pone.0195029

Birnbaum, M. H., & Mellers, B. A. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology*, *45*(4), 792–804. doi: 10.1037/0022-3514.45.4.792

Chapman, G. B., & Liu, J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgment and Decision Making*, *4*(1), 34–40.

Cohen, A. L., & Staub, A. (2015). Within-subject consistency and between-subject variability in Bayesian reasoning strategies. *Cognitive Psychology*, *81*, 26–47. doi: 10.1016/j.cogpsych .2015.08.001

Dohmen, T., Falk, A., Huffman, D., Felix, M., & Sunde, U. (2009). *The non-use of Bayes rule: Representative evidence on bounded rationality* (Tech. Rep.). Maastricht University, Research Centre for Education and the Labour Market. (ROA-RM-2009/1, February 2009)

Engelmann, D., & Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review*, *94*(4), 857–869.

Galesic, M., Gigerenzer, G., & Straubinger, N. (2009). Natural frequencies help older adults and people with low numeracy to evaluate medical screening tests. *Medical Decision Making*, *29*(3), 368–371. doi: 10.1177/0272989X08329463

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704. doi: 10.1037/0033-295X .102.4.684

Güth, W., Schmidt, C., & Sutter, M. (2007). Bargaining outside the lab–a newspaper experiment of a three-person ultimatum game. *The Economic Journal*, *117*(518), 449–469.

Hafenbrädl, S., & Hoffrage, U. (2015). Toward an ecological analysis of bayesian inferences: how task characteristics influence responses. *Frontiers in psychology*, *6*, 939–939. doi: 10.3389/fpsyg.2015.00939

Hayes, B. K., Ngo, J., Hawkins, G. E., & Newell, B. R. (2018). Causal explanation improves judgment under uncertainty, but rarely in a Bayesian way. *Memory & Cognition*, *46*(1), 112–131. doi: 10.3758/s13421-017-0750-z

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, *73*(5), 538–40. doi: 10.1097/00001888-199805000-00024

Hoffrage, U., Hafenbrädl, S., & Bouquet, C. (2015). Natural frequencies facilitate diagnostic inferences of managers. *Frontiers in Psychology*, *6*, 642. doi: 10.3389/fpsyg.2015.00642

Hoffrage, U., Lindsey, S., & Hertwig, R. (2000). Communicating statistical information. *Science*, *290*(5500), 2261–2262. doi: 10.1126/science.290.5500.2261

Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review*, *116*(4), 856–874. doi: 10.1037/a0016979

Juslin, P., Nilsson, H., Winman, A., & Lindskog, M. (2011). Reducing cognitive biases in probabilistic reasoning by the use of logarithm formats. *Cognition*, *120*(2), 248–267. doi: 10.1016/j.cognition.2011.05.004

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. doi: 10.1006/obhd.2000.2895

McKenzie, C. R. (1994). The accuracy of intuitive judgment strategies: Covariation assessment and Bayesian inference. *Cognitive Psychology*, *26*(3), 209–239. doi: 10.1006/cogp.1994.1007

McNair, S., & Feeney, A. (2015). Whose statistical reasoning is facilitated by a causal structure intervention? *Psychonomic Bulletin & Review*, *22*(1), 258–264. doi: 10.3758/s13423-014-0645-y

Mellers, B. A., & McGraw, A. P. (1999). How to improve Bayesian reasoning: Comment on Gigerenzer and Hoffrage (1995). *Psychological Review*, *106*(2), 417–424. doi: 10.1037/0033-295X.106.2.417

Nadanovsky, P., dos Santos, A. P. P., Lira-Junior, R., & de Oliveira, B. H. (2018). Clinical accuracy data presented as natural frequencies improve dentists' caries diagnostic inference: Evidence from a randomized controlled trial. *The Journal of the American Dental Association*, *149*(1), 18–24. doi: 10.1016/j.adaj.2017.08.006

Nozick, R. (1969). Newcombs problem and two principles of choice. In *Essays in honor of Carl G. Hempel* (pp. 114–146). Springer.

Pighin, S., Gonzalez, M., Savadori, L., & Girotto, V. (2016). Natural frequencies do not foster public understanding of medical test results. *Medical Decision Making*, *36*(6), 686–691. doi: 10.1177/0272989x16640785

Vallée-Tourangeau, G., Abadie, M., & Vallée-Tourangeau, F. (2015). Interactivity fosters Bayesian reasoning without instruction. *Journal of Experimental Psychology: General*, *144*(3), 581–603. doi: 10.1037/a0039161

Weber, P., Binder, K., & Krauss, S. (2018). Why can only 24% solve bayesian reasoning problems in natural frequencies: Frequency phobia in spite of probability blindness. *Frontiers in Psychology*, *9*, 1833. doi: doi:10.3389/fpsyg.2018.01833

Woike, J. K., & Kanngiesser, P. (2019). Most people keep their word rather than their money. *Open Mind*, *3*, 68–88.

Zhu, L., & Gigerenzer, G. (2006). Children can solve Bayesian problems: The role of representation in mental computation. *Cognition*, *98*(3), 287–308. doi: 10.1016/j.cognition.2004.12.003

Zizzo, D. J. (2003). Money burning and rank egalitarianism with random dictators. *Economics Letters*, *81*(2), 263–266.