Multi-Task Reinforcement Learning in Humans: Supplementary Information

Momchil S. Tomov[1,2,*,†], **Eric Schulz**[3,4,*,†], **and Samuel J. Gershman**[2,4,5]

[1]Program in Neuroscience, Harvard Medical School, Boston, MA 02115, USA
[2]Center for Brain Science, Harvard University, Cambridge, MA 02138, USA
[3]Max Planck Institute for Biological Cybernetics, Tübingen, 72012, Germany
[4]Department of Psychology, Harvard University, Cambridge, MA 02138, USA
[5]Center for Brains, Minds and Machines
[*]Corresponding authors: Momchil S. Tomov (mtomov@g.harvard.edu) and Eric Schulz (eric.schulz@tuebingen.mpg.de).
[†]Contributed equally.

ABSTRACT

Supplementary Information for “Multi-Task Reinforcement Learning in Humans”.

Supplementary Information

Model Specifications

We formally specify all models below. Even though it is possible to change specific parts of the models’ implementations, all of our predicted effects are largely independent of implementational details.

For all models, we evaluated performance on the test trial using a softmax policy based on the learned Q-values:

$$\pi_{\text{test}}(a|s) \propto e^{\beta Q(s,a)}, \tag{1}$$

where $\pi_{\text{test}}(a|s)$ is the probability of choosing action a in state s on the test trial, $\beta = 10$ is the inverse temperature and Q is the Q-value function computed during training by the respective model. Since we were primarily interested in comparing models based on asymptotic performance, this ensured that the models were placed on an equal footing, even though they were trained using different methods and training policies. Softmax is also consistent with human behavior[?].

Model-free learning

We trained a model-free agent on the same training tasks and environment as the participants. We used tabular Q-learning:[?] after taking action a and transitioning from state s to state s', the Q-value for state-action pair (s,a) was updated as:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r(s') + \gamma \max_{a'} Q(s',a') - Q(s,a)], \tag{2}$$

where $r(s) = \phi(s)^\top \mathbf{w}$ is the reward in state s for training task $\mathbf{w}$, $\alpha = 0.1$ is the learning rate, and $\gamma = 0.99$ is the discount factor. Values were mapped to choices using an ϵ-greedy exploration strategy with $\epsilon = 0.1$:

$$\pi_{\text{train}}(a|s) = \begin{cases} 1-\epsilon & \text{if } a = \operatorname{argmax}_a Q(s,a) \\ \epsilon/2 & \text{otherwise} \end{cases} \tag{3}$$

Q-values were initialized randomly between 0 and 0.00001 to break ties initially. Each training task was encountered for 200 training episodes (trials). Note that this approach ignores the features ϕ and task weights $\mathbf{w}$ and only considers states, actions, and the final reward.

Model-based learning

We assume the model-based agent has perfect knowledge of the environment. We computed the Q-values for each task $\mathbf{w}$ separately using value iteration:? the value $V(s)$ of each state s was updated according to:

$$V(s) \leftarrow \max_a Q(s,a), \tag{4}$$

where

$$Q(s,a) = \sum_{s'} p(s'|s,a)[r(s) + \gamma V(s')], \tag{5}$$

where $r(s) = \phi(s)^\top \mathbf{w}$ is the reward in state s for task $\mathbf{w}$, $p(s'|s,a)$ is the probability of transitioning from state s to state s' after taking action a (note that in our deterministic setup, $p(s'|s,a) = 0$ or 1), and $\gamma = 0.99$ is the discount factor. This update was repeated for all states until convergence, defined as the largest update being less than 0.01. Note that, unlike model-free learning, value iteration was also applied to the test task.

Universal value function approximators

One way to train a UVFA online involves using a deep Q-learning network. Since we were interested in asymptotic performance, we instead trained it in a supervised way.? For each training task, we computed the Q-values using value iteration as described in the model-based section. We then trained a 3-layer feedforward function fitting neural network (fitnet in MATLAB) with 10 hidden units to predict the Q-values for each training task. The transitions between states were deterministic, $Q(s,a) = V(s')$, where $p(s'|s,a) = 1$, so for inputs to the network we used $(\mathbf{s}, \mathbf{w})$-tuples, where $\mathbf{s}$ is a one-hot encoding of the state and $\mathbf{w}$ is the task weights vector. Each training tuple and the corresponding Q-value were repeated 100 times. The Q-values for the test task were generated as the predictions of the network for $(\mathbf{s}, \mathbf{w}_{\text{test}})$.

Successor features and generalized policy improvement

We assume the agent learns a policy and its corresponding successor features for each training task.? While these could be computed online using temporal difference learning, we were primarily interested in asymptotic performance, so we computed the optimal policy for each training task using value iteration, as described in model-based section. We then computed the successor features for each policy π using dynamic programming by iteratively applying the Bellman equation until convergence (maximum change of less than 0.01):

$$\psi^\pi(s) \leftarrow \phi(s) + \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)\psi^\pi(s') \tag{6}$$

We computed the Q-values on the test task $\mathbf{w}$ using generalized policy improvement. This involved iterating over the policies and computing the expected reward of the test task $\mathbf{w}$, using the successor features for each policy π:

$$Q^\pi_{\mathbf{w}}(s,a) = \psi^\pi(s,a)^\top \mathbf{w} \tag{7}$$

The Q-value for each state-action pair was then chosen based on the policy that will perform best on the test task $\mathbf{w}$:

$$Q^{\text{max}}(s,a) = \max_\pi Q^\pi_{\mathbf{w}}(s,a) \tag{8}$$

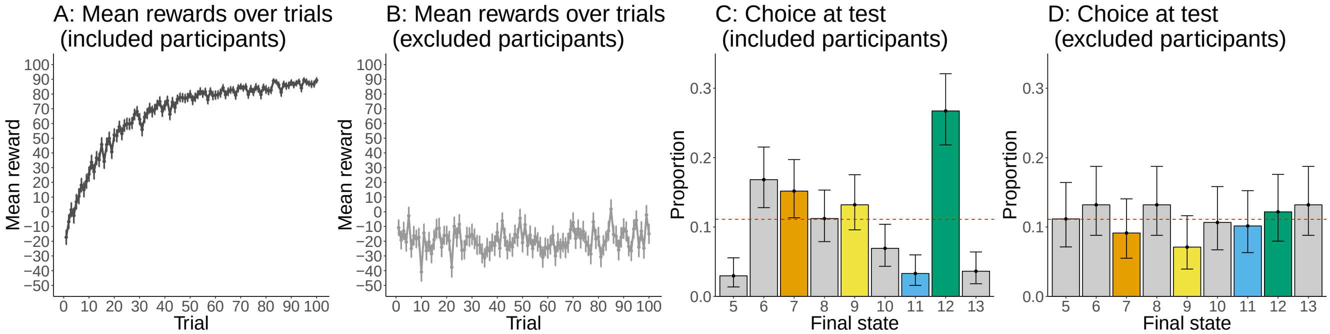
The values $Q^{\text{max}}(s,a)$ were then entered into the softmax choice rule (Eq. 1). Notice that the only difference between this approach and the original SF&GPI formulation is that the latter deterministically takes the argmax action. Using softmax instead merely adds noise to choices (and converges to argmax when $\beta \to \infty$) without altering the main prediction, while ensuring that all models are compared on an equal footing.

Effect of exclusion criterion

We assess the effect of exclusion criteria onto our main finding of participants choosing the SF&GPI state more frequently than what was expected under the chance level of $p = 1/9$. First, we assess the proportions of people choosing the SF&GPI state without any exclusion of participants. Doing so, we find that 40 out 226 participants choose state 12 in Experiment 1

($\hat{p} = 0.18$, exact binomial test: $p = .003$, $BF = 12.2$), 48 out of 202 participants chose state 12 in Experiment 2 ($\hat{p} = 0.24$, exact binomial test: $p = 4.22 \times 10^{-7}$, $BF = 38408$), and 42 out of 200 participants chose state 12 in Experiment 3 ($\hat{p} = 0.21$, exact binomial test: $p = 6.21 \times 10^{-6}$, $BF = 434$). Next, we assess what happens to our results if we exclude participants who achieved an average reward below the median during the training trials. Doing so, we find that 21 out of 113 participants chose state 12 in Experiment 1 ($\hat{p} = 0.19$, exact binomial test: $p = .016$, $BF = 3.3$), 30 out of 106 participants chose state 12 in Experiment 2 ($\hat{p} = 0.28$, exact binomial test: $p = 1.17 \times 10^{-5}$, $BF = 1376$) and 31 out of 96 participants chose state 12 in Experiment 3 ($\hat{p} = 0.32$, exact binomial test: $p = 2.42 \times 10^{-8}$, $BF = 3.39 \times 10^{5}$). We therefore conclude that our main effect is largely independent of the chosen method of participant exclusion.

Comparing included with excluded participants



Supplementary Figure 1. Differences between included and excluded participants in Experiment 4. **A:** Included participants' ($N_{\text{incl}} = 303$) mean rewards over trials. **B:** Excluded participants' ($N_{\text{excl}} = 197$) mean rewards over trials. **C:** Included participants' choices on the 101st trial. **D:** Excluded participants choices on the 101st trial. Error bars represent the standard error of the mean in A and B, and 95% confidence intervals in C and D.

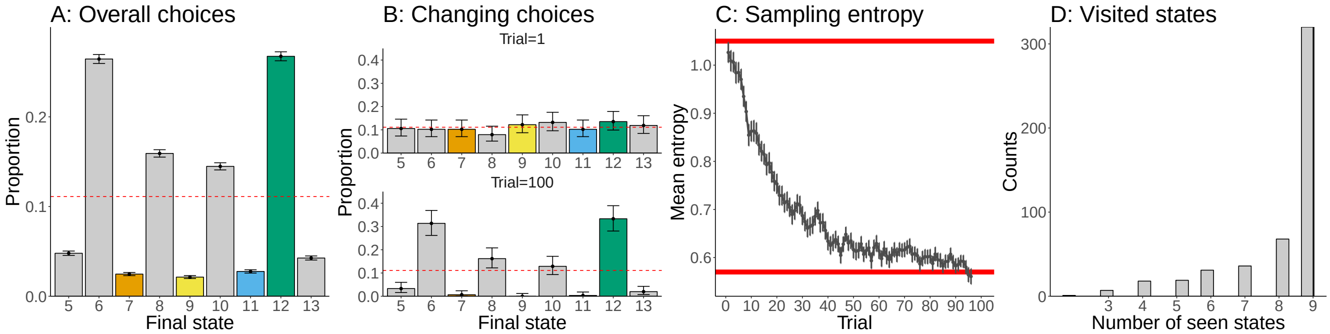
One of the underlying assumptions when excluding larger proportions of participants is that their behavior was vastly different from the included participants' behavior. To verify this assumption, we analyzed both included and excluded participants' behavior in Experiments 4. Specifically, we looked at participants' learning curves during training as well as their choice behavior on the 101st trial (Supplementary Figure 1). Participants who accumulated an average reward of higher than 0 showed clear signs of learning across trials, leading to a positive correlation between trial number and average reward (Supplementary Figure 1A; $r(98) = 0.85$, $p < .001$, $BF = 2.9 \times 10^{25}$). Participants who accumulated an average reward of less than 0 showed no signs of learning across trials and their average rewards did not improve over trials (Supplementary Figure 1B; $r(98) = 0.12$, $p = .23$, $BF = 0.45$). Participants' choices also differed vastly between the two groups. Whereas the included participants showed the choice pattern described in Experiment 4 (see **??**B and Supplementary Figure 1C), the excluded participants did not choose any of the final states more frequently than chance ($\chi^2(8) = 6.25$, $p = .62$, $BF = 1.5$). These results show that the excluded participants did not learn training tasks and therefore ended up performing at chance on the final test trial, further justifying our exclusion criterion.

Further analysis of training data

We also further analyzed participants' behavior during the learning phase of Experiment 4, in order to assess whether the main assumptions of the winning model (SF&GPI) were indeed fulfilled as indicated by participants' choices on the 100 training task trials (Supplementary Figure 2).

The first assumption was that participants sampled both state 6 and state 12 in equal proportions during the learning trials. We wanted to ensure that participants had no bias to sample state 12 more frequently to begin with, which we expected since both states were rewarded equally often. Analyzing participants' choices during the learning trials, we found no difference between the frequency of choosing state 6 ($\hat{p} = 0.26$) and the frequency of choosing state 12 ($\hat{p} = 0.27$, $\chi^2(1) = 0.51$, $p = 0.237$, $BF = 0.03$). We also calculated the difference between the frequency of choosing state 6 and choosing state 12 for each participant, and compared the resulting mean difference per participant against 0 using a within-subjects test. This test also revealed no difference in frequencies between the two states ($t(302) = 0.28$, $p = 0.78$, $d = 0.02$, $BF = 0.07$). Thus, the assumption that participants engaged in both policies, one leading to state 6 and one leading to state 12, in equal proportion was met (Supplementary Figure 2A).

The second assumption was that participants had found and converged to the policies leading to state 6 and state 12 by the end of the learning trials. We therefore analyzed participants' choices on the 1st and the 100th trial. Whereas participants



Supplementary Figure 2. Analysis of training data in Experiment 4. **A:** Proportion of choices over all 100 training trials. **B:** Proportion of choices on the 1st trial (top) and the 101st trial (bottom). **C:** Sampling entropy over a window of 4 consecutive choices. Red lines mark theoretical lines of sampling between all options (top line) and sampling between two options (bottom line). **D:** Proportion of total number of seen states over all participants. Error bars represent the 95% confidence interval for A and B and the standard error of the mean for C.

did not choose any of the states more frequently than chance on the 1st trial ($\chi^2(8) = 6.77$, $p = .561$, $BF = 0.46$), they chose both state 6 ($\hat{p} = 0.33$, binomial test: $p = 2.2 \times 10^{-16}$, $BF = 1.17 \times 10^{18}$) and state 12 ($\hat{p} = 0.33$, binomial test: $p = 2.2 \times 10^{-16}$, $BF = 3.1 \times 10^{21}$) more frequently than chance on the 100th trial. As before, there was also no significant difference in the frequencies of choosing state 6 and state 12 on the 100th trial ($\chi^2(1) = 0.51$, $p = 0.237$, $BF = 0.03$). Thus, the assumption that participants started out exploring different states but ended up converging on the policies leading to state 6 or 12 was met (Supplementary Figure 2B).

The third assumption was that participants had stopped exploring and converged on exploiting the best final states on the 100th trial. We therefore calculated each participants sampling entropy, measured by Shannon's entropy, over a moving window of 4 consecutive samples and then averaged those entropies over all participants for each trial (Supplementary Figure 2C). This showed that participants started out exploring the states as evenly as a random sampler would at the beginning of the experiment but then quickly converged to the same entropy as a two-state sampler. Since we know from the analysis reported in Supplementary Figure 2B that participants most frequently sampled states 6 and 12 on the last trial, the analysis of their sampling entropies further corroborated the fact that they had indeed converged on exploiting the policies that lead to those states. Thus, the assumption of convergence in participants' sampling behavior was met.

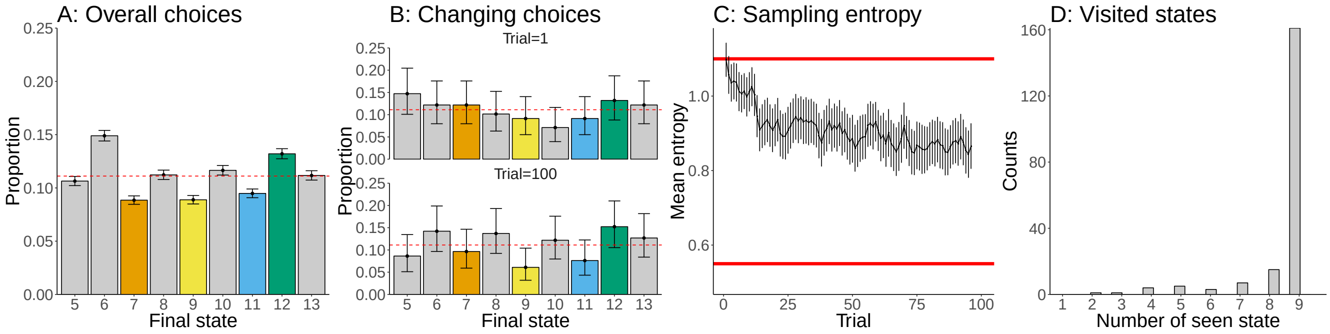
The final assumption was that most participants had indeed seen all final states or at least were not driven by a lack of knowledge about state 7, which was the best possible choice on the 101st trial and also predicted by a model-based planner. We therefore analyzed how many unique states each participant had seen over all 100 learning trials (Supplementary Figure 2D). This showed that the majority of participants, 159 out of 303, saw all 9 states at least once. Only looking at the 4 states which were predicted by the different models, we found that 231 of 303 of all participants had seen all of the four predicted states. 265 of 303 participants had sampled state 7 at least once during the first 100 trials. Importantly, whether or not a participant had previously visited state 7 was not predictive of choosing that state on the 101st trial ($BF = 0.21$). Additionally, how often someone had visited state 7 was not predictive of whether or not state 7 was chosen on the 101st trial ($BF = 0.06$). Finally, the total number of unique states a participant had visited during the first 100 trials was also not predictive of choosing state 7 on the 101st trial ($BF = 0.48$). Thus, we can conclude that our finding of participants preferably choosing state 12 (which was predicted by the SF&GPI model) on the 101st trial was not influenced by not having seen enough alternative states or never having seen the state predicted by a model-based planner.

Further analysis of excluded participants' training data

We analyzed the excluded participants' behavior during the learning phase of Experiment 4, in order to assess whether these participants did indeed exhibit learning behavior that was vastly different from the included participants' behavior (Supplementary Figure 3).

First, we assessed the excluded participants' proportion of overall choices during the 100 training trials (Supplementary Figure 3A). Similar to the included participants, there was no difference in the proportion of choosing state 6 $\hat{p} = 0.15$) and state 12 ($\hat{p} = 0.13$, $\chi^2(1) = 1.47$, $p = 0.236$, $BF = 0.03$). However, the proportions of choosing either of these states was considerably smaller than for the included subjects. Thus, the excluded participants did not learn about the two underlying policies as well as the included participants.

Next, we analyzed whether the excluded participants had converged to the policies leading to state 6 and state 12 by the end of the learning trials. We therefore analyzed participants' choices on the 1st and the 100th trial as before. The excluded



Supplementary Figure 3. Analysis of excluded participants' training data in Experiment 4. **A:** Proportion of choices over all 100 training trials. **B:** Proportion of choices on the 1st trial (top) and the 101st trial (bottom). **C:** Sampling entropy over a window of 4 consecutive choices. Red lines mark theoretical lines of sampling between all options (top line) and sampling between two options (bottom line). **D:** Proportion of total number of seen states over all participants. Error bars represent the 95% confidence interval for A and B and the standard error of the mean for C.

participants did not choose any of the states more frequently than chance on the 1st trial ($\chi^2(8) = 8.08, p = .426, BF = 0.73$) and on the 100th trial ($\chi^2(8) = 14.66, p = .066, BF = 1.00$). Thus, the excluded participants did not converge on the policies leading to state 6 or 12 (Supplementary Figure 3B).

The third analysis concerned participants' sampling entropy, measured by Shannon's entropy, over a moving window of 4 consecutive samples per participant and then averaged over all participants for each trial (Supplementary Figure 3C). This showed that participants started out exploring the states as evenly as a random sampler would at the beginning of the experiment and *never* converged to the same entropy as a two-state sampler. Even more, the excluded participants' entropy over time did not look like adding many more learning trials would have been enough for them to converge on the optimal policy, because their average entropy seemingly plateaued after about 20 trials. Thus, the assumption of convergence in participants' sampling behavior was not met in the excluded participants' data.

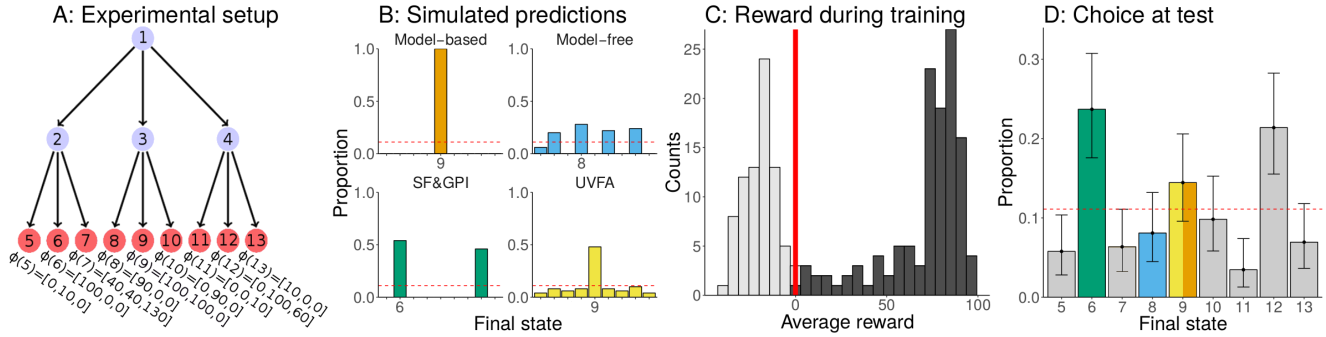
Finally, we also assessed how many unique states the excluded participants visited during the learning trials. 161 out of 197 participants visited each state at least once, and all of the excluded participants visited each of the states predicted by the four different models at least once during the learning trials. Therefore, the difference between excluded and included participants could not be explained by the number of uniquely visited states.

Taking these results together, we conclude that the excluded participants did not converge to the two optimal policies during the learning trials.

Further testing predictions of UVFAs

One of our motivating examples concerned a situation in which an agent was trained on two weights independently and then was tested on a task where both weights were present. Since this situation offers another test case for multi-task reinforcement learning, and in particular one in which UVFAs could match with participants' behavior, we additionally assessed people's behavior in this scenario in another experiment. Both the training weights and the features were the same as in Experiment 3 and 4 (Supplementary Figure 4A). However, the test weights for the 101st trial were set to $\mathrm{w}_{\text{test}} = [1, 1, 0]$. In this scenario, both model-based planning and UVFAs predict that participants will sample state 9 on the 101st trial (Supplementary Figure 4B). Whereas model-free planning does not fully converge on any state, the most likely state to be sampled according to this model is state 8. Finally, the SF&GPI model predicts that participants will choose either state 6 or state 12, with a small preference for state 6.

We ran this experiment by recruiting 201 participants (92 females, mean±s.d. age: 34.98±9.1 years) from Mechanical Turk following the same procedure as before. Similarly to all of our previous experiments, the distribution of mean rewards during the first 100 trials was bimodal and we therefore removed participants with a mean reward of lower than 0. This led to the removal of 78 participants. Out of the remaining 123 participants, 34 participants chose state 6 on the 101st trial, which was significantly above the chance level of $p = 1/9$ ($\hat{p} = 0.28$, one-sided exact binomial test: $p = 3.65 \times 10^{-7}$, $BF = 29738$). The second most frequently chosen option was state 12, which was chosen by 29 participants, again significantly above chance ($\hat{p} = 0.28$, one-sided exact binomial test: $p = 6.74 \times 10^{-5}$, $BF = 29738$). Only 20 out of 123 participants chose state 9, which was predicted by both the UVFA model and by model-based planning algorithms. The proportion of participants choosing this state was not significantly different from chance ($\hat{p} = 0.16$, one-sided exact binomial test: $p = .05$, $BF = 1.08$). Thus, even in an experiment where both the UVFA and the model-based planning algorithm predicted state 9, only the states predicted by the SF&GPI model were chosen significantly more frequently than chance.



Supplementary Figure 4. Overview and results of additional experiment. **A:** Experimental setup. Participants ($N = 201$) are trained on the set of weights $\mathbf{w}_{\text{train}} = \{[1, -1, 0], [-1, 1, 0], [1, -2, 0], [-2, 1, 0]\}$ and tested on the weights $\mathbf{w}_{\text{test}} = [1, 1, 0]$. The features for each final state are shown below the tree. **B:** Predictions of the different models. Predictions were derived by simulating models given the training weights and then registering their decisions given the weights of the test trial. This simulation was repeated 100 times for each model and the proportions of choosing the different target states was tracked. **C:** Distribution of average reward obtained by participants during the training trials. Participants were split into a group that accumulated less than 0 points (gray, $N_{\text{excl}} = 78$) and a group that accumulated more than 0 points (black, $N_{\text{incl}} = 123$), which we analyzed further. Red vertical line marks the threshold of 0. **D:** Participants' ($N_{\text{incl}} = 123$) choices given the new weights $\mathbf{w}_{\text{test}}$ on the test trial. Choices are colored by the simulated model predictions. Error bars show the 95% confidence interval of the mean based on an exact binomial test. Dashed line indicates chance responding.